1-1-1975

An empirical comparison of criterion-referenced data collected by mastery testing versus repeated item-examinee sampling.

Peter Edward Schriber
University of Massachusetts Amherst

AN EMPIRICAL COMPARISON OF CRITERION-REFERENCED DATA COLLECTED BY MASTERY TESTING VERSUS REPEATED ITEM-EXAMINEE SAMPLING

A Dissertation Presented

By

Peter Edward Schriber

Submitted to the Graduate School of the University of Massachusetts in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

January, 1975

Major Subject: Educational Evaluation

AN EMPIRICAL COMPARISON OF CRITERION-REFERENCED DATA COLLECTED BY MASTERY TESTING VERSUS REPEATED ITEM-EXAMINEE SAMPLING

A Dissertation

By

PETER EDWARD SCHRIBER

Approved as to style and content by:

William Phillip Gorth, Chairman of Committee

Ronald K. Hambleton, Member

Richard D. Konicek, Member

Louis Fischer, Acting Dean, School of Education

Doctor of Education

January 1975

ACKNOWLEDGEMENTS

This dissertation was three and one half years in the making during which time it changed form several times and was reduced in scope to the present four questions. I owe a large debt of gratitude to my wife, Cathy, who bore the brunt of my long hours of writing, rewriting and computer work. My chairman, Dr. William Gorth, gave me many moments of time and much advice throughout. Roy Williams served as a technical advisor for statistics and I owe much to him.

A very special thanks is to be extended to Mr. John Easter, Assistant for Research, Sequoia Union High School District, Redwood City, California, for giving permission for his district to host my study and for providing aid in recruiting teachers and students to participate.

Drs. Ronald K. Hambleton, Richard D. Konicek, and LeVerne J. Thelen of the School of Education provided valuable counsel and assistance.

A word of caution is to be added for those who will read this dissertation. The introductory section which summarizes thinking in criterion-referenced testing is current through 1972 only. New thinking and research has greatly expanded the field in the 1973-1974 interim. Appendix C is a brief summary of the research and thought of the latter two years as it relates to the focus of this study. Also, the results of this study apply specifically to the item-examinee sampling design which is currently referenced as the Standard Comprehensive Achievement Monitoring (CAM) design and, thus, do not apply to all matrix sampling designs in general.

ABSTRACT

AN EMPIRICAL COMPARISON OF CRITERION-REFERENCED DATA COLLECTED BY MASTERY TESTING VERSUS REPEATED ITEM-EXAMINEE SAMPLING (January, 1975)

Peter E. Schriber, B.S., M.S., State University College, New Paltz, New York
Ed.D. University of Massachusetts
Directed by: Dr. William Phillip Gorth

This study focuses on the use of criterion-referenced testing in the classroom. Two systematic procedures for constructing and using criterion-referenced tests were compared as to their ability to provide data useful to the teacher for making decisions about individual student achievement. One procedure was mastery testing, the posttest measurement of individual student performance. The other type of testing compared was Comprehensive Achievement Monitoring (CAM). The basic CAM design employs item-examinee sampling whereby a series of randomly parallel test forms would measure all objectives for a semester with each objective measured by several test items across forms. In sum, mastery tests focus on individual student mastery and CAM tests on group data.

The purpose of the study was to compare CAM testing to mastery testing to determine if CAM does complement mastery testing, i.e., do the two modes of testing provide data of practical use to the teacher and are the two types of testing actually furnishing different types of information? The key

research questions posed were: (1) Are items which appear on both types of tests answered differently (i.e., have different item difficulties)? and (2) Are early CAM and mastery test scores useful in predicting student success in the course?

The sample was 256 ninth and tenth grade students taking general biology. There were six CAM and four mastery test administrations given during one semester. Students were arranged into eight testing groups with each group taking six different CAM forms in a prescribed sequence.

Major results were:

(1) Encountering an item first on a CAM test (before instruction) and later on a mastery test had no effect on student performance on the item on the mastery test. However, students performed significantly more poorly on an item when it was encountered after instruction on a CAM test than when it was encountered on the mastery test.

(2) The first mastery test score and the score of a standardized verbal ability measure were much better predictors of the global measures of student progress and of final semester grade than the early CAM scores. A second analysis of the usefulness of several prediction equations to the teacher for making decisions about individual student progress showed the standardized verbal

test score and first mastery test score to be equally useful for predicting success. The least useful were early CAM scores.

The results of this study have implications for the classroom teacher. The analyses show the strength of teacher-constructed criterion-referenced tests in providing much useful data for classroom management. Additionally, the last set of analyses respond directly to a statement made by Hively, et al.,[1] (1973) in which the diagnosis of a student's learning may be useful in predicting "his behavior (in a non-statistical, inductive fashion) in natural situations which have some properties in common with the test items" within the content domain (underlining added). Statistical prediction was achieved by this study to properties related to the content domain. Early mastery tests were shown to be useful in predicting final course success giving the teacher the opportunity to meet individual needs in instruction. The results indicate that the early differentiation of students on future course success may be better done (and done well) by early short-term mastery posttests than by tests which sample content of a large portion of the course.

---

[1]W. Hively, G. Maxwell, G. Rabehl, D. Sension, & S. Lundin, Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project, CSE Monograph Series in Evaluation, No. 1, Los Angeles: Center for the Study of Evaluation, University of California, 1973, page 15.

TABLE OF CONTENTS

TABLE OF TABLES

CHAPTER I

INTRODUCTION

1.1 Background

Within the last 15 years there have been widespread movements toward the individualization of instruction in education and toward the formulation of curricula based on explicitly stated objectives for the outcomes of instruction. Several now well-known models for individualized instruction have been developed. The acronyms from three of these models have entered the jargon of education signifying categories of instructional approaches. The three models are "computer-assisted instruction" (CAI) (Suppes, 1966; Atkinson, 1968; Atkinson and Wilson, 1970), "program for learning in accordance with needs" (PLAN) (Flanagan, 1967, 1969), and "individually prescribed instruction" (IPI) (Glaser, 1968).

The emphasis on objective-based curricula has resulted in a movement back to the use of behaviorally stated learning outcomes popular in this country, particularly in the areas of mathematics and grammar, during the 1930s and 1940s. Thirty years ago Dodd (1943) suggested a procedure for "operationalizing" statements of goals for learning so that outcomes could be measured through the observation of

the learner's behavior. Now with the strong emphasis on meeting the needs of the individual through instruction, the renewed interest in designing curricula which lend themselves to the measurement of individual progress on absolute versus relative (i.e., normative) standards has led to a proliferation of methods for identifying, constructing, using and selecting behaviorally stated objectives (Bloom, Hastings, and Madaus, 1971).

With the advent of the "taxonomy of educational objectives" (cognitive domain: Bloom, 1956; affective domain: Krathwohl, Bloom, and Masia, 1964; and the parallel taxonomy of Gagne, 1965) the general educational community was presented with a means of structuring objectives within a curriculum. This structuring was the ingredient generally lacking from the objectifying schemes of the 1930s and 1940s. Not only was behavior and content of educational objectives required to be specified, but the behavior specified in the objective was placed within a hierarchial framework. Added to this were adaptations of the computer to educational applications (e.g., computer-assisted instruction) and the combination of the computer and operationalized objectives in programmed instruction. In all probability Mager's (1962) book entitled Preparing Objectives for Programmed Instruction was read by more people not involved in programmed instruction than by people who were involved.

The demand for behaviorally stated, or what is now generally termed "performance-based" curricula, has been ex-

pressed by the numerous objective and test item banks coming into prominence such as the Instructional Objective Exchange (IOX) (Popham, 1971) and the New York State Education Department's large and well-organized bank for the State's System for Program Planning, Evaluation and Development (SPPED) (O'Reilly, 1973). Many textbooks, following in the footsteps of Mager, Bloom, and others, are describing procedures for constructing behavioral objectives and using them in the classroom (e.g., Mager, 1972; Sund and Picard, 1972; Tanner, 1972; Vargas, 1972).

All these changes in emphasis on the modes of teaching and of constructing and using curricula have not left the area of measurement untouched. Measurement to determine whether students have achieved specifically stated learning outcomes requires that the learning of the objective be demonstrated by the performance of the behaviors stated in the objective. Thus, the individual's performance is measured in terms of absolute criteria rather than the relative criteria of his rank on a test or content domain with a specified group. Tests which collect such data are termed "criterion-referenced" tests, a term which appears to have first entered the educational measurement literature in an article by Glaser (1963).

The present study considered the application of criterion-referenced testing to the classroom with emphasis on gathering data for program management rather than for assessment of individual student progress. Specifically, the

purpose of this research was to document several of the strengths and limitations of criterion-referenced testing for instructional decision making by contrasting two uses of the testing. One use was gathering data for individual student guidance. This type of testing has been called "mastery testing" or "unit testing" and will be more fully defined in Section 1.1.3. The second use of criterion-referenced testing was gathering data for instructional program decisions. The focus of the data is on group measurement rather than individual measurement which is the focus of the data in mastery testing. A particular application of this use of criterion-referenced testing is the major component of Comprehensive Achievement Monitoring (CAM) (Allen, 1968, Allen and Gorth, 1971). In 1971 the CAM technique was a formative evaluation model wherein a set of course-representative test forms are constructed and administered in a longitudinal testing design to gather data useful in program evaluation. (Since 1971 the CAM model has been expanded to include criterion-referenced mastery testing designs as well as the longitudinal item-examinee testing designs [see for example O'Reilly, Gorth, and Pinsky, 1973]. The present study was conducted when the acronym CAM represented the longitudinal testing design only.) The CAM model will be more fully described in Section 1.1.2. It was selected for contrast with mastery testing for student guidance because: (1) it is a well-defined, documented, and piloted technique used for formative program evaluation and (2) it has been advertized

as being capable of supplying supplementary data useful for program decisions which complement data from mastery testing and other data-gathering techniques whose principal purpose is individual student assessment (Allen, 1968; Allen and Gorth, 1971).

1.1.1 The instructional model sensitive to individual differences. Maintaining an instructional environment sensitive to individual differences and the necessity of making program decisions requires that instruction and learning function within a cybernetic system (Merrill, 1968; Glaser and Nitko, 1971). This system is characterized by the capacity to govern itself and provides for information to be collected on student performance useful in instructional decision making for both individuals and groups in a cyclical manner. It is a closed system in that it is self-contained and self-regenerative (Merrill, 1968). The system operates when vital information is collected at the appropriate intervals and instructional decisions are made when necessary and based on the information gathered. Such decisions include the periodic re-assessment and refinement of the system.

The components necessary for a functional cybernetic instructional system have been identified by Glaser and Nitko (1971). In addition to the periodic feedback for system improvement, there are five additional essential components:

1. Instructional objectives are specified in terms of observable student behavior.

2. Before a student begins a segment of instruction, a diagnosis of his initial learning capabilities is made.

3. The student is placed in instruction according to the diagnosis of his state of learning.

4. Student performance is monitored and assessed continuously as learning occurs.

5. Instruction proceeds in a fashion determined by the available resources, assessment of performance, and standards of learning competence.

The research performed for this present study addressed components 2 and 4 as well as the cybernetic feedback framework. The chief type of data useful for instructional decision making is criterion-referenced data (Glaser, 1963; Popham and Husek, 1969; Glaser and Nitko, 1971). The two types of criterion-referenced testing compared in this study relate directly to the need for two foci of achievement data necessary for program management, the individual and the group.

1.1.2 Comprehensive Achievement and Monitoring (CAM) testing. CAM is a formative evaluation model which employs criterion-referenced testing to gather longitudinal data for a group of students on all objectives being tested in a program including student achievement, curriculum, and instructional treatment. The CAM technique functions as a systematic procedure for continuous performance assessment. Its design has the following components:

1. A curriculum composed of behaviorally stated objectives and test items referenced directly to specific objectives.

2. Item sampling in the form of a series of randomly parallel test forms where each test form contains items representative of the program.

3. Tests administered longitudinally across the course.

4. Analysis of test data and reporting of results by computer shortly after each test occasion.

5. Use of data by program managers, evaluators, teachers, and students for decisions regarding instructional treatment, curriculum, and the CAM design being used.

To implement CAM, the student population is divided into "student schedule groups" (SSGs) for the purpose of testing. The term "schedule" refers to the unique order in which each SSG is administered the CAM test forms. An SSG is a sample of students representative (i.e., a simple or stratified random sample) of the population. All members of a SSG take the same test form at a test administration. Each SSG is equal in size to every other. No SSG takes the same test form as another at a given test administration and, typically, no SSG takes the same form twice.

The result of administering a set of randomly parallel test forms repeatedly during a program to the student groups is data on all tested instructional objectives at every administration. Therefore, data are available for instructional decision making at multiple points during a program.

CAM was developed to provide: (1) an efficient method for measuring learning and (2) an effective feedback of results to students, teachers, and school administrators. Over the past five years (1967-1971) CAM has been used with nearly

50,000 students in 10 states (Allen and Gorth, 1971). The trend data it returns for decision making are based on three time periods for a given instructional objective: preinstruction, immediate postinstruction and intervals a month or later following instruction.

The course-representative nature of the test forms employed provides that every student begins taking tests that are mostly pretest in content. Gradually, as more instruction occurs, the tests provide both pretest and posttest data, pretest-posttest-later retention data, and finally posttest and later retention data only. The object is to produce group-trend data on all course objectives gathered at frequent intervals across all students which is immediately useful for instructional decision making usually on a biweekly or monthly basis.

The elements in the CAM design are not new. What is new is the use of what Campbell and Stanley (1963) call a "quasi-experimental" data-gathering design in classroom evaluation with criterion-referenced testing. The data-gathering design is a type of time-series design wherein a series of measurements or observations are conducted over time on an individual or group. The purpose of the design is to determine the effect of a given intervention occurring at a specific point in time. The diagram of the design is as follows with "O" designating an observation and "X" the point of intervention (e.g., instructional treatment):

$$O_1\ O_2\ O_3\ O_4\ O_5\ XO_6\ O_7\ O_8\ O_9\ O_{10}$$

Campbell and Stanley (1963) report a quasi-experimental, time-series design which closely approximates the CAM data-gathering design. The design is termed a "counterbalanced design" in which precision of experimental control is sought by entering all subjects into all treatments. One method of achieving this, which CAM does, is to use a Latin-square arrangement. Thus, using as an example a counterbalanced design of a series of four CAM test forms and four student schedule groups (SSGs), the CAM design would appear as follows (where the subscripts designate CAM test forms):

TABLE 1.1.1

The CAM Design as a Time-Series Counterbalanced Design

Student Schedule Group (SSG)	Test Occasion			
	1	2	3	4
1	X_1O	X_3O	X_4O	X_2O
2	X_2O	X_1O	X_3O	X_4O
3	X_3O	X_4O	X_2O	X_1O
4	X_4O	X_2O	X_1O	X_3O

Note.--X = point of intervention
O = an observation

Counterbalancing is present in that:

1. Each SSG is representative of the population and equal in size.
2. Each SSG takes the four CAM test forms (1, 2, 3, and 4) at different occasions.
3. Each test form is course representative.
4. Due to conditions 1, 2, and 3 above, change in performance in any column of O's compared to any other column would be interpreted as a change in student learning.

This counterbalancing can be summarized by pointing out that each combination of variables in Table 1.1.1 (groups, occasions and testing treatments) occurs equally often (thus, a Latin-square design) (Campbell and Stanley, 1963). Again, the idea is an application of a previously developed and used experimental design dating at least to McCall (1923) when it was then termed a "rotation experiment."

1.1.3 Unit testing. Unit testing is used in this study synonymously with the term "mastery testing." Unit tests are designed to measure the performance of individual students on one or more objectives or concepts. The purpose of the test is to determine individual mastery over the set of objectives or concepts involved. This is accomplished by having one or more items on the test for each objective or concept tested. The test generally is used as a pretest or posttest for the particular segment of study represented by the test content. Typically, unit tests do not overlap in the objectives they measure and thus provide either a single

pretest or posttest estimate of achievement, or a pretest-posttest pair of achievement estimates of individual performance. Thus, the data are useful for individual student guidance. Such tests, therefore, can be used either as diagnostic instruments or as measures of achievement.

The term "unit testing" was used in this study to denote that the tests measure performance on units of objectives which were the basic components of the course in the study.

1.1.4 Program management and instructional decision making. The purpose of this study was to contrast the two types of criterion-referenced testing: the CAM testing, which focused on obtaining group data and unit testing, which focused on obtaining individual student data. Important in such a comparison is whether or not the two techniques complement each other in the data they supply for classroom and program management.

Educational measurement provides the information necessary for making decisions concerning the development, operation, evaluation, and refinement of the instructional process (Glaser and Nitko, 1971). Often the chief source of data for making such decisions has come from teacher-constructed pretests and posttests on segments of the course under study. The general purpose of this testing has been to manage the learning environment of the individual student. Program management is sometimes equated with the guidance of the individ-

ual student through testing or classwork which supplied data considered relative only to that individual's performance. However, other decisions are necessary in program management in addition to focusing merely on guidance of the individual through a program (Merrill, 1968; Glaser and Nitko, 1971). Even to make sound decision on individual student achievement, the rationale behind this research study was that data other than individual student pretest and posttest mastery data are necessary. To identify these data needs, an examination of the types of instructional decisions is fruitful. O'Reilly and Gorth (1972) classify instructional decisions into four general categories:

1. Type 1: Decisions for guidance of the individual student through a program.
2. Type 2: Ongoing program decisions, i.e., those involving the development, refinement, and justification of the instructional design which includes the instructional mode and materials and minor curricular modifications and refinement.
3. Type 3: Curriculum development, re-organization, and revision beyond minor refinements to existing curricular packages.
4. Type 4: Comparative product evaluation where one instructional treatment is compared to another in terms of effectiveness.

Typically, teachers usually make decisions of Types 1 and 2 while specialists deal with Types 3 and 4. As mentioned before, for Type 1 decisions traditional classroom testing in the form of mastery testing provides the data base. Type 2 decisions are ongoing program decisions most often

made by the teacher and include curricular refinements such as resequencing instructional objectives for presentation and modifying individual objectives, instructional approaches and materials. Large-scale reformulations of curricula are Type 3 decisions and are generally the task of curriculum development teams. Comparative product evaluation (Type 4 decisions) involves the comparison of instructional treatments and is the concern of program managers and research specialists.

Program management typically involves the continual making of decision Types 1 and 2. Group performance data, which summarize the group's performance in the program, are necessary for putting an individual's performance into perspective. Questions relating to ongoing program decisions are relevant to individual guidance and include: Are several students having the same failures or successes? Is instruction misleading to more than this one student? Is instruction guiding or aiding the students as expected? and Is the cause of failure faulty student performance, faulty specification of an objective or faulty teaching? Both group and individual student data are necessary for making decisions relevant to answering these questions and to selecting appropriate instructional alternatives to improve the instructional situation.

CAM is designed to provide data for decision Types 2, 3, and 4; unit or mastery testing is designed to provide data for decision Type 1.

1.2 Statement of the Problem

The interrelationships of CAM and unit testing within a single setting are the prime focus of the research problem of this report. Unit testing provides individual achievement data while CAM provides group data for the three time phases of pre-instruction, post-instruction, and later retention. The question arises of whether or not this is a true description of the relationship of the two criterion-referenced testing strategies. The answer to this question lies in an investigation of the issues concerning the interrelationship of the two.

One issue is whether or not the data gathered from the two types of criterion-referenced testing serve different purposes. CAM gathers data primarily useful for decision making about groups; unit testing gathers data primarily useful for decision making about individual students (or groups at one point in time for a segment of the program). A second issue concerns the actual ability of the CAM technique to provide information which supplements individual assessment data for instructional decision making.

The questions investigated in this report focus on these issues in a manner that provides results directly applicable to the classroom situation and chief user of the data, the teacher. Each of the following topics is the nucleus for a research question:

1. Differences among CAM test-form difficulties and the effect of differences on data analysis,
2. The effect on test item difficulty of an item which appears on both a longitudinal (i.e., CAM) and unit (i.e., mastery test,
3. The association of CAM and unit test scores and the possible overlap in information that is provided by the two, and
4. The usefulness of early CAM test scores in predicting global measures of individual student progress.

Little formal study has been done on the use and effects of CAM testing (two notable exceptions are Pinsky and Gorth, 1969a, and 1969b). The present endeavor was a much more detailed approach than any previous with several analytical approaches not tried with CAM data previously.

1.3 Purposes of the Research

Four research questions were formulated to address the topics listed under the statement of the problem for research.

Question (1). Are there differences among CAM test-form difficulties at each test administration and across test administrations?

"Test-form difficulty" is defined as the mean test score of a group obtained on a given test form at a test administration. To provide data for decision-making purposes, CAM test forms must be reliable and valid indicators of performance for each test administration. With criterion-

referenced testing a basic assumption is that the test items are valid measures of the objectives. Thus, several items for one objective should measure equally the examinee's performance of the objective and, therefore, should be equally difficult. Test forms which are equally difficult are necessary but not sufficient for test reliability and validity.

If test forms are not equally difficult, one cannot identify individual and group growth across time. If test forms fluctuate in relative difficulty at each test administration, any attempt to correct for test difficulty will vary in effectiveness from test administration to test administration. This would make trend plotting most difficult and tenuous. The expected change in test-form difficulty would be that all forms should become easier at the same rate across test administrations and this would be the result of student learning.

Another consideration in longitudinal, criterion-referenced testing is the repeated use of the same test items. If nonequivalence in test-form difficulty were discovered in test piloting or early in the usage of the tests, corrective measures can be taken to replace faulty items so that the test user (most often the classroom teacher) will not repeatedly give a poor test and receive misleading test results.

Question (2). a. What is the effect on item difficulty of students having encountered an item on a CAM test form prior to encountering it on a unit posttest?

b. Are there differences between the difficulty of post-instruction items first encountered on CAM test forms and the difficulty of the same items first encountered on unit posttests?

"Item difficulty" is defined as the proportion of subjects responding correctly out of the total group of subjects. Question 2 was desired to probe the effect on item difficulty of subjects encountering an item twice: once on a CAM test form and later on the unit posttest. The results have implications for test security and for distortions in the test scores of individual students.

One of the implications for test security involves the manner in which results are reported to the student. The student is not given back his corrected test form; rather, he is given a report of his performance on the objectives to which the items were referenced: whether he answered correctly, incorrectly, or left the items blank. The student does not receive the items themselves; rather, a general practice is to suggest similar items which he may work on or, to give him the basic objectives and sources of available materials for further study. Thus, hopefully a source of test bias is averted.

If an item appears on both tests, some students will have seen the item previously either on the CAM or unit test form depending on the particular student schedule group. Second, an effect may be present due to the nature of the two

kinds of criterion-referenced tests (the unit test being all posttest and the CAM test having pretest, posttest, and later retention components). Students may perform differently on CAM posttest items than on identical unit posttest items. There may be differences in motivation for taking the two tests. Students may view the CAM test as a means of setting up baseline data for further diagnostic approaches. A test configureation of items representing three phases of instruction may affect performance on identical items appearing on unit tests.

In terms of decision making by the teacher, it is important to know if similar decisions can be made based on seemingly similar types of data (i.e., posttest data) from CAM and unit testing. Thus, this question bears on whether or not there is duplication in data gathered by the two testing strategies which could be used for the same decision.

Of particular interest to test users would be the finding that a test item could be used on both types of tests with little or no confounding effects. If such a result were found for a variety of subject areas and student age levels, a case could be made for the use of repeated test items on CAM tests.

Question (3). What are the correlates of various ability measures with several global measures of the success of individual students in a course?

This question concerns the relationships between global measures of success for individuals and test scores which may be potential predictors of these measures. Comparing unit test measures and CAM test measures relates to the possible overlap of data provided by the two. Similar correlations between the unit test scores and CAM test scores and the global success measures would suggest data overlap. Question 3 is also a prerequisite to Question 4 which is based on the relationships found among the global measures of course success and potential predictors of success, and whose results provide the prediction possibilities of the most likely potential predictors. Global measures include:

(1) sum of unit posttest scores,

(2) sum of CAM scores,

(3) sum of CAM pre-instruction scores,

(4) sum of CAM instruction-completed scores (instruction-completed scores are the subscores of CAM tests that include items from objectives on which instruction has been completed),

(5) teacher's final course grades, and

(6) CAM test scores of the final CAM test administration.

Test score variables selected as potential predictors were selected on the basis of their early availability in the course. The test scores selected were early CAM scores, the first unit posttest score, and a SCAT II verbal subtest score. SCAT II (Cooperative School and College Ability Tests, Series 2) scores were available for most students. These tests were

selected to measure the general abilities of students and contain verbal and mathematical ability subtests. A single total score for the verbal subtest was available to the teacher before the beginning of the course.

In addition to identifying overlap in unit data gathering and CAM data gathering, the correlations among global success measures and potential predictors are useful to the teacher in identifying students who are heading for trouble in a course.

> Question (4). Do early unit test scores and CAM test scores predict global measures of individual student progress and final course grade?

Question 4 was designed to compare the value of the selected predictor variables for prediction purposes, i.e., their usefulness for making decisions on possible global success and failure of students based on early indicators. Such information is, of course, useful to the test user for planning for individual student instruction as well as group instruction.

Summary. These four questions bear directly on the establishment of the value of CAM as a complement to well-constructed and systematically constructed unit tests. CAM is a complement if it can provide the instructional decision maker reliable data not available from unit testing.

1.4 Limitations

Several limitations on the testing techniques used for comparison in this study need to be stated. First, the unit tests were not as much different from the CAM tests as would be best for results of a study such as this. Many of the objectives were tested by only one or two test items providing inconclusive results as to determination of individual mastery.

There are also limitations in the CAM technique. One is the necessity of having a group of students working with the same instructional objectives (not necessarily at the same time) in order to produce trend data. This poses some problems for CAM design when gathering data in programs with individually prescribed or paced instruction. Following this limitation is the need for a student population of a certain size for a specified number of objectives in order to be able to test performance on all the objectives with test forms of reasonable lengths. Modifying test forms when it is found desirable to alter the CAM design during a program can be a problem. One possible solution is computer-generated test forms (see Gorth and Grayson, 1971; Schriber and Gorth, 1971).

Another problem is that of the effect of repeated pretesting on the same objectives with the same students. This effect has only been minimally studied (see Gorth, Allen, Popejoy, and Stroud, 1968). Another possible limitation is

the problem of test security. This includes the possibility that students may remember items from a previous test or may give the items or their answers to other students who have not yet taken the particular test forms involved.

CHAPTER II

RESEARCH DESIGN

The research questions posed in this study called for a research design employing both CAM testing and unit posttesting in an instructional setting and required that both criterion-referenced testing strategies share the same objectives and some portion of identical test items. The questions required the CAM tests and unit posttests to test the same objectives with both different and identical test items.

Mandatory for a study of this kind that uses CAM tests and unit posttests is that the participants, both teachers and students, should have had previous experience with the two types of criterion-referenced testing and not be negatively disposed to using them. Also to obtain wholehearted participation from teachers, objectives and test items need to have been used and refined previously by the teachers themselves. The teachers should be comfortable with the curriculum and the test items and should want to use them. In addition, it is desirable to have the teachers involved in the selection of objectives and test items and in the construction of tests for both CAM and unit testing. When teach-

ers are involved to the extent of creating the curriculum and testing instruments, their commitment to them becomes more integral to their goals and tasks. This desirable situation existed with the teachers of this study.

It would be a difficult task to find another situation so well suited to the purposes of this study as the one used; in addition to previous experience with CAM by teacher and student alike, both types of criterion-referenced testing were present for an existing course and curriculum. Favorable also was that most students completed the entire testing treatment which lasted a full semester. Planning such a study from the start for a given student sample necessitated numerous decisions in curriculum development, test construction, and training of students and teachers. The present study accomplished all of the above. In addition, the students and teachers had used both CAM and unit testing concurrently during the previous school year.

2.1 Sample

The study population consisted of 299 ninth and tenth grade biology students (256 of whom completed the entire test treatment) taking General Biology at the Menlo-Atherton High School of the Sequoia Union High School District, Redwood City, California. The 43 students who had not completed the test treatment had either taken forms in an incorrect sequence, repeated a form, or had missed from one to nine tests. The scores of these 43 students were omitted from the study. The

population was composed of 11 class sections (see Table 2.1.1) taught by four different teachers. Even though the experimental situation of the study was not highly controlled, the total of 256 of 299 students with complete test results over a semester's duration and 10 tests is 85.6 percent of the total or a relatively high total given normal absenteeism and attrition of students.

TABLE 2.1.1

Number of Students and Class Sections by Teacher

Teacher Identification Number	Number of Class Sections	Number of Students	Number of Students in Study[a]
16	1	29	25
51	4	96	76
57	2	53	39
70	4	121	116

[a]Students who completed the testing treatment.

2.2 Course Description

Instructional treatment was essentially the same for each class section and all students moved through instruction in a group-paced manner. The curriculum closely followed a single biology text. Laboratory experience was required and accounted for a third of a student's final semester grade. All four teachers previously had used the same curriculum

with CAM and unit testing.

2.3 Content of Tests

A total of 89 objectives were represented on the CAM and unit tests with objectives weighted with from one to 10 items each. All items were of the four-response, multiple-choice type. CAM tests measured performance on 84 objectives, two of which were not measured by unit testing. Unit tests measured performance on 87 objectives, five of which were not measured by CAM testing. Objectives were divided into 16 units containing from one to 16 objectives each. The weighting of objectives with widely varying numbers of items and units with varying numbers of objectives made the comparison of results among objectives and among units more difficult than if weighting had been approximately equal. This weighting is the chief uncontrolled variable of the study.

There were eight CAM test forms containing a total of 272 items. Nine of these appear on two forms so that there are actually only 263 different items. Each CAM form was representative of the objectives for the entire semester.

Four unit posttests were constructed with each one testing a different segment of the semester. The four tests contained a total of 150 items with none repeated. CAM and unit test form composition is displayed in Table 2.1.2 and the relationship of objectives to lessons and units is displayed in Table 2.1.3. "Lesson" is a term designating a set of objectives which, after instruction was completed on the

TABLE 2.1.2

Test Form Composition: Number of Test Items on Each Test Form from Each Item Category

Item Category	Test Form Number											
	51	52	53	54	55	56	57	58	61	62	63	64
On both CAM and unit tests	13	11	9	10	10	10	8	12	20	22	22	18
On CAM tests only, but objective represented on unit tests also	20	23	24	23	23	22	25	21				
On CAM tests only, but objective not represented on CAM tests also	1	0	1	1	1	2	1	1				
On unit tests only, but objective represented on CAM tests also									16	18	16	12
On unit tests only, but objective not represented on CAM tests									4	0	2	0
Total number of items per test	34	34	34	34	34	34	34	34	40	40	40	30

TABLE 2.1.3

Structure of Curriculum by Unit, Objective, Lesson, and Items per Unit on CAM Tests and Items per Unit on Unit Posttests

Unit ID No.	No. of Objectives in Unit	Objective ID Nos.	Lesson Containing Unit	Lesson Completion Date	Items per Unit on CAM Tests	Items per Unit on Unit Tests
11	2	1103-1104	1	9/20/71	2	2
12	6	1201-1202, 1204-1207	1	9/20/71	3	6
13	7	1301-1307	1	9/20/71	12	12
14	1	1402	2	10/01/71	2	2
15	2	1501-1502	2	10/01/71	4	2
16	4	1601-1604	2	10/01/71	9	7
21	5	2101-2105	3	10/13/71	16	9
22	5	2201, 2203-2306	4	10/20/71	16	5
23	8	2301-2304 2305-2308	5 6	10/25/71 10/28/71	32	9
24	5	2401-2405	7	11/23/71	16	6
25	16	2501-2505 2506-2510 2511-2516	8 9 10	11/23/71 11/23/71 11/23/71	48	20
26	8	2601-2608	11	12/07/71	16	24
27	6	2701-2703 2704, 2706-2707	12	12/10/71	32	16
28	5	2801-2805	14	1/11/72	16	10
29	5	2901-2905	15	1/28/72	16	11
30	4	3001-3002 3003-3004	16 17	1/28/72 1/28/72	32	9
Total	89				272	150

set, signaled the time for posttesting.

As shown in Table 2.1.2, items repeated on both CAM and unit tests account for nearly a third of each CAM test and one-half of each unit posttest. Table 2.1.3 shows plainly the weighting of units with unequal numbers of objectives and the consequent weighting of the units by unequal numbers of items on both the CAM and unit tests. Lesson 1, which includes Units 11, 12, and 13, contained review and prerequisite objectives for the course and received the lightest weightings.

Other variables considered in the study were: (1) SCAT II verbal subtest scores which were available for 208 of the 256 students and (2) teacher's final semester grade available for all 256 students in the form of letter grades: A, B, C, D, and F.

2.4 Testing Schedule

There were a total of 10 test administrations of which six were CAM and four were unit test administrations. The duration of the testing was one semester, September 1971 through January 1972. The six CAM test administrations provided each student with six of eight CAM test forms over the semester. All four of the different unit posttests were administered, each at a different test administration. The testing schedule is presented in Table 2.1.4. From Table 2.1.4 it can be seen that the interval between CAM test administrations varied from 16 to 40 days and was generally

near 20 days. The unit test administrations also occurred at unequal intervals, the reason being that the lessons into which the course material was divided were of varying sizes and required different amounts of time for completion. Also several intervals contained more than one lesson. Appendix A contains a description of each test form by item, objective, unit, lesson, and date of completion of each lesson.

TABLE 2.1.4

CAM and Unit Test Administrations by Date (September 1971 to January 1972) and Number of Items per Test Form

Entry	Test Administration									
	1	2	3	4	5	6	7	8	9[a]	10[a]
Date	9/15	10/5	10/14	10/29	11/24	12/8	12/17	1/12	1/28	1/28
Type of test	CAM	CAM	Unit	CAM	Unit	CAM	Unit	CAM	Unit	CAM
Number of items/ test form	34	34	40	34	40	34	40	34	30	34

[a]Test administrations 9 and 10 occurred at the same time and the resulting composite scores were used as the final semester test.

Even though each objective was not tested by an equal number of test items (a situation which would have simplified dealing with the research questions), a very large amount of useful data was provided for the purposes of the study.

CHAPTER III

RESULTS AND DISCUSSIONS

3.1 Question (1)

Are there differences among CAM test-form difficulties at each test administration and across administrations?

Question 1 consists of two parts. The first part asks whether CAM test-form difficulties are equal. The second asks if the relative ordering of test-form difficulties is significantly changed at each test administration since fluctuating test-form difficulties provide trend data that are difficult or impossible to rely on.

3.1.1 Analysis of the Equivalence of CAM Test-Form Difficulties

3.1.1.1 Design. The design of the analysis for equivalence of CAM test-form difficulty includes the calculation of the descriptive statistics of mean total test score and standard deviation of the test scores across students taking each test form at each test administration. This allows an examination of where some of the differences occurred. The second part of the design is an analysis of variance to

determine whether the differences observed among test-form difficulties at each test administration are significant.

The equivalence of the test-form difficulty of the eight test forms at each of the six test administrations separately was investigated by one-way analysis of variance. The general model for the completely randomized one-factor design was used under the assumption that subjects were randomly assigned to each of the student schedule groups. The model is as follows:

$$Y_{ij} = M + a_j + e_{ij} \qquad (1 - 1)$$

where Y_{ij} = score of subject i on test form j,

M = mean of the population

a_j = effect of test form j, and

e_{ij} = the deviation due to variability of the score of subject i on test form j from the jth treatment population mean (i.e., error variance of the score of subject i on test form j).

The null hypothesis for each test administration was that the test forms would be equally difficult; that is, their effects on student scores would be the same at a given test administration regardless of test form. This may be stated as:

$$H_o : a_1 = a_2 = \ldots = a_8,$$

where a_j is the effect of the jth test form. The alternative hypothesis is that the test forms are not equivalent in difficulty, that is:

$$H_1 : a_1, a_2, \ldots, a_8 \text{ are not equal.}$$

3.1.1.2 Sample. The sample consisted of 259 students.

3.1.1.3 Data. The data were the total test scores of each student for each of the six CAM test administrations.

3.1.1.4 Results. The mean and standard deviation of the total test scores of the students taking each CAM test form at each administration are presented in Table 3.1.1. The means and standard deviations for: (1) each test form across all test administrations and (2) each test administration across all test forms, were computed and also entered in Table 3.1.1. The number of observations per cell varied from 29 to 37 because the eight student schedule groups were unequal in size.

The row means of Table 3.1.1 are the mean test-form difficulties for the CAM test forms across the six test administrations. The mean test-form difficulties varied from 15.9 to 19.5; therefore, the forms were not equivalent in difficulty. The desired finding would have been equivalent test-form difficulty. Test Forms 52, 53, and 54 were most difficult with mean scores of 15.9, 16.0, and 16.6. Another four test forms were somewhat easier but nearly equivalent in difficulty to one another with mean scores of 17.5, 17.6, 17.7, and 18.0. One form was clearly the easiest of the eight with a mean score of 19.5. Mean variation of test scores per form (as measured by the mean of standard deviations for the six test administrations for each form) across

TABLE 3.1.1

Mean and Standard Deviation of CAM Test-Form Difficulty at Each Test Administration

Test Form	Test Administration (and Date) 1 (9/15)	2 (10/5)	3 (10/29)	4 (12/8)	5 (1/12)	6 (1/28)	Row Mean
51	(29)	(35)	(30)	(34)	(31)	(33)	(192)
	11.8	13.5	17.0	18.1	21.1	23.5	17.5
	3.1	3.9	3.9	5.1	4.6	5.5	4.5
52	(30)	(29)	(37)	(33)	(29)	(35)	(193)
	11.6	12.4	15.8	17.5	17.7	19.9	16.0
	2.9	4.1	4.0	4.7	4.0	5.1	4.3
53	(37)	(34)	(31)	(35)	(30)	(29)	(196)
	11.8	12.5	15.2	18.2	18.9	20.4	15.9
	3.1	3.4	3.7	3.9	4.1	4.5	3.9
54	(34)	(31)	(33)	(31)	(30)	(36)	(195)
	12.3	12.3	15.5	17.9	18.4	22.9	16.6
	2.9	4.1	3.8	3.5	3.9	5.2	4.0
55	(31)	(33)	(30)	(29)	(36)	(34)	(193)
	13.1	14.2	16.2	18.2	21.2	22.2	17.6
	3.2	4.4	4.6	3.1	4.6	5.0	4.3
56	(35)	(30)	(29)	(31)	(33)	(29)	(187)
	13.5	14.7	15.7	19.8	21.5	23.0	18.0
	3.1	2.7	3.7	3.8	5.2	4.5	4.0
57	(33)	(29)	(35)	(37)	(33)	(31)	(198)
	14.0	16.1	17.5	21.4	23.1	25.0	19.5
	4.2	3.6	3.4	5.2	4.2	5.3	4.5
58	(29)	(37)	(34)	(29)	(35)	(30)	(194)
	13.6	15.1	15.5	18.8	20.6	23.2	17.7
	3.2	3.5	4.0	4.7	5.0	4.3	4.2
Column Mean*	(258)	(258)	(259)	(259)	(257)	(259)	(1550)
	12.7	13.9	16.1	18.8	20.4	22.3	17.4
	3.3	3.8	3.9	4.4	4.6	5.0	4.2

Note.--For each cell in the body of this table, the three numbers represent:

(1) the number of observations per cell (in parenthesis),
(2) mean of the total test score,
(3) standard deviation of these scores.

* All totals were not 259 because three students included were discovered to have taken only five of the six CAM tests. Their scores were omitted from the analyses of Questions 2, 3, and 4.

test administrations was fairly uniform for the eight test forms with a range of 3.9 to 4.5.

The analyses of variance to determine whether the test-form difficulties were significantly different from one another at each test administration are summarized in Table 3.1.2. The BMDO1V computer program (UCLA, 1964a) was used to calculate the six analyses of variance.

The null hypothesis was rejected for each test administration with the exception of the fourth administration. Therefore, the alternative hypothesis that test-form difficulty is not equal at each test administration with the exception of the fourth is accepted. However, a significant F-ratio does not necessarily mean that all test-form difficulties were different from one another at a given test administration. In this particular set of analyses of variance the difficulty of two test forms is observably different from the others and contributes most to the significant F-ratios. In Table 3.1.1 it can be seen that Test Form 52 was consistently more difficult than the other seven forms with exceptions at Test Administration 2 where it was slightly easier than Test Form 53 and Test Administration 4 where all forms were found not to differ significantly in difficulty. Test Form 57 was consistently the least difficult form at all six test administrations.

TABLE 3.1.2

Analysis of Variance Among CAM Test Forms at Each Test Administration

Test Administration	F-Ratio	Level of Significance[a]
1	2.470	$p < .025$
2	4.201	$p < .001$
4	1.362	--
6	2.924	$p < .01$
8	4.612	$p < .001$
10	3.542	$p < .005$

[a]Level of significance given only if $p < .05$.

3.1.2 Analysis of Relative Ordering of CAM Test-Form Difficulties

3.1.2.1 Design. This analysis was concerned with determining whether or not there was a CAM test form by test-administration interaction. The focus is not just on whether or not the relative ordering of CAM test-form difficulties changes across test administrations, but on whether or not any significant change is due to subject interaction effects. Therefore, rather than simply calculating a Spearman-Rho correlation, an analysis of variance of test-form difficulties across test administrations was calculated.

The CAM design is a repeated-measures approach with the student groups, in this instance, student schedule groups, taking the CAM test forms in a particular sequence. There

are three main effect variables: test form, test administration, and student schedule group. Therefore, an effect of test form at test administration in student schedule group may produce two and three-way interactions. The model which includes these variables is:

$$Y_{ijkr} = M + f_i + a_j + g_k + (fa)_{ij} + (fg)_{ik} + (ag)_{ik} + (fag)_{ijk} + e_{ijkr}, \qquad (1 - 2)$$

where: Y_{ijkr} = the score of individual r of student schedule group k at test administration j on test form i,

M = the mean of the population,

f_i = the effect of form i,

a_j = the effect of administration j,

g_k = the effect of group k,

$(fa)_{ij}$ = the effect of test form i at administration j,

$(fg)_{ik}$ = the effect of test form i in student schedule group k,

$(fag)_{ijk}$ = the effect of form i at administration

e_{ijkr} = the error variance of the score of individual r of student schedule group k at administration j on form i.

The letter "r" was used to denote an individual student because all students in a student schedule group repeat the treatment in an identical manner in the basic CAM longitudinal testing design.

The parameters of Model 1 - 2 are not all estimable. In a CAM design no two student schedule groups take the same

form at the same test administration and all forms are administered at each test administration. Therefore, the three two-way interactions in the above model are confounded with their complementary main effects because test forms and student schedule group combinations change at each test administration. However, with a repeated measures design, more power is gained which unconfounds the main effects from the interaction terms. (Myers, 1966)

Therefore, the appropriate analysis of variance model does not have two-way interaction terms. The model is:

$$Y_{ijkr} = M + f_i + a_j + g_k + (fag)_{ijk} + e_{ijkr} \qquad (1 - 3)$$

with the terms being synonymous with those of Model 1 - 2.

To ascertain the presence of the three-way interaction effect of test form (i) at test administration (j) in student schedule group (k), the Least Squares and Maximum Likelihood General Purpose Program (LSMLGP) (Harvey, 1968) was selected because it handles the analysis of variance with an unequal number of observations in cells. The numbers of observations in the cells are unequal because the student schedule groups vary in size. The significance of the three-way interaction must be estimated in two steps because of computer program limitations. In Step One an analysis of variance is performed using the simple additive model where the error term (or residual) contains any interaction effect variance as well as error variance attributable to individual subject differences. The model for Step One is:

$$Y_{ijkr} = M + f_i + a_j + g_k + \text{residual}, \qquad (1 - 4)$$

where the variables are the same as those for Model 1 - 2 with the exception of "residual" which is the total of all interaction effect variances and error variance.

In Step Two, a one-way analysis of variance was calculated for each of the 48 levels consisting of each of the eight student schedule groups at each of the six test administrations (i.e., each of the 48 cells in the CAM design). The residual sum of squares from Step Two is substracted from the residual sum of squares from Step One. The difference is the sum of squares for the three-way interaction, $(fag)_{ij}$. The subtraction removes all two-way interaction effects. Thus, the necessary F-ratio for a three-way interaction can be computed.

The null hypotheses for the two-step analysis of variance were:

(1) no interaction effect of test form, test administration, and student schedule group was present,

$$H_{o_1} : (fag)_{ijk} = 0,$$

(2) in addition, there was no main effect of test form,

$$H_{o_2} : f_i = 0 \text{ and } (fag)_{ijk} = 0, \text{ and,}$$

(3) further, there was no main effect of student schedule group,

$$H_{o_3} : f_i = 0,\ g_k = 0,\ \text{and}\ (fag)_{ijk} = 0.$$

The alternative hypotheses were that any or all of the effects in each of the null hypotheses were not equal to zero.

3.1.2.2 Sample. An N of 217 was used rather than the total N of 256. There are several reasons for the use of a restricted sample of subjects. All of the students taught by one of the four teachers in the study were omitted because: (1) the data of the class were poorly collected with 25 percent of the students missing data and (2) evidence from an inte-view that the teacher did not use CAM results from several test administrations. Of a total of 53 students in the two sections of this teacher originally included in this study (N = 299), the scores of 14 were discarded (due to missing data or to the taking of test forms in incorrect sequence). This proportion of rejection was larger than that of any of the other three teachers. With the exclusion of the scores of the students of all four teachers have missing data or incorrect testing sequence, the total N was 256. However, many of the remaining 39 students of the teacher with the largest percentage of rejection also showed irregularities in CAM scores. Twenty-six scored higher on the first CAM test than on the second and a total of 21 scored higher on the first test than on the third. Therefore, these 39 were also dropped for the two-step analysis of variance leaving an N of 217.

3.1.2.3 Data. The data consisted of the raw CAM total test scores for the 217 students for five of the six CAM test administrations. The scores of the sixth (final) CAM administration were used as part of a final semester examination given in conjunction with and including a final unit posttest. The scores of the final CAM administration were omitted because of the possible introduction of bias due to the testing situation being different from those of the earlier tests.

3.1.2.4 Results. Table 3.1.3 contains the summary of Step Two in the two-step approach to the analysis of variance. It shows that there were significant main effects of test administration of test form (both $p < .001$) but not of student schedule group. It would be expected that test administrations would show differences in test scores (in this instance a consistent trend for scores to increase). Also it was known before the analysis that test forms differed from one another in difficulty. A crucial condition for the comparison of test-form difficulties is that the student schedule groups should be equivalent in test performance. This was supported with an F-ratio of 0.95 for main effect of student schedule group. Thus, null hypotheses H_{o_2} and H_{o_3} were rejected due to the main effect of test form.

The major point of the analysis of variance summarized in Table 3.1.3 was to determine whether or not the three-way interaction of test administration/test form/student schedule

group was significant. It would be significant if the test forms varied in relative difficulty from one another at each administration and with a particular student schedule group. In this study this would be true when any of the 48 cells of the analysis of variance matrix differed significantly from any other in regard to the pattern that the main effects show across the data. This interaction was found not be be present with an F of 0.86. Therefore, the null hypothesis H_{o_2} that there was no three-way interaction cannot be rejected.

TABLE 3.1.3

Results of the Analysis of Variance for 217 Students Over the First Five CAM Test Administrations to Determine if CAM Test Forms Varied in Relative Difficulty from One Another Across Test Administrations

Source of Variance	df	Sum of Squares	Mean Squares	F-Ratio
Total	1085	26235		
Test administration (TA)	4	8651	2162.8	140.64**
Test form (TF)	7	957	136.8	8.89**
Student schedule group (SSG)	7	102	14.6	0.95
TA/TF/SSG	21	278	13.3	0.86
Residual	1045	16070	15.4	

** $p < .001$

Omitting the sixth CAM test administration may not have been necessary because a cursory glance at Table 3.1.1 shows the sixth administration to be similar in pattern to the other five administrations.

3.1.3 Discussion

Question 1 considered the hypothesis that the eight CAM test forms are equivalent in difficulty both at each test administration and across test administrations. If differences were found at each test administration among the eight forms, the relative positions of the test forms as to their difficulties at each administration would have to be examined to determine if group profiles were possible.

Implications of the findings of the two parts of Question 1 for the use of CAM designs in general follow.

First, the importance of having student schedule groups which perform equivalently on CAM tests is highlighted by these analyses. Equivalence depends on how representative each group is of the entire population. Otherwise, test-form difficulty becomes dependent on each student schedule group. The analysis of variance to determine if there is change in relative test-form difficulty among test administrations becomes very hard to interpret if the test forms actually are equivalent but do not appear so due to nonequivalence of student schedule groups.

Second, for users of CAM data the knowledge of test-

form difficulty and standard deviation would add to a more complete interpretation of group scores at least at the total test score level and perhaps at the unit and objective levels. This knowledge may even be useful when working at the level of the total test score of an individual student.

Third, if student schedule groups are equivalent, it tentatively appears that test forms do not change in relative difficulty over time relative to each other. Thus, if CAM scores are summed for individual students, the sums are comparable because they are based on the same tests, taken in different sequences, and free from shifts in relative position of test-form difficulty. Variance in relative position of test-form difficulty would make comparison between scores of the same or different individuals impossible or highly tentative. Thus, reliability of the test forms would be low.

Fourth, given no difference in relative test difficulty across administrations, estimates of test-form difficulties can be made over time with small samples as long as the groups are representative of the larger population. This would be valuable in field testing CAM test forms before implementation in a project.

3.1.4 Limitations of the Analyses

A basic problem for the analyses performed for Question 1 was that the CAM design fails to meet the criteria for univariate analyses of variance for determining if differences

in relative test-form difficulties occur across test administrations. The problem is the failure to meet the assumption of the randomness of error present in the six test scores for each student. For a given individual his six CAM scores may have a consistent component which is part of the variance of the score. For example, a high-achieving student would be expected to score consistently higher on each successive test. The low-achieving student probably would not be expected to do the same, or to do so within narrow limits. Therefore, to obtain a more accurate estimation of the interaction effect of test form, administration, and student group, a multivariate analysis of variance model must be developed (see Swaminathan, 1972).

3.2 Question (2)

(A) What is the effect on item difficulty of students having encountered an item on a CAM test prior to encountering it on a unit posttest?

(B) Are there differences between the difficulty of posttest items first encountered on CAM test forms and the difficulty of the same items first encountered on unit posttests?

Both questions consider the effects of using some identical test items to carry out concurrently the two strategies for criterion-referenced testing, CAM testing and unit

posttesting. For Question 2A the difficulty of items (expressed as proportion of students answering each item correctly) on a unit posttest were compared for those students who had encountered the same items before instruction on the associated objective on a CAM test form and those students who had no prior experience with the items. The situation is one of using test items measuring students' performance on an objective by one strategy and being concerned about the effect on the measurement of the same objective later by the other strategy. If no effects are detected for identical items, probably no effects would be obtained for nonidentical items. Question 2B is considered after the report for the analysis of Question 2A.

3.2.1 Analysis of Question 2 (A)

3.2.1.1 Design. The item difficulty of 10 randomly selected items from Unit 26 by student schedule group and by test administration were presented in tables so that the raw data for comparison would be available for inspection. The Wilcoxon matched-pairs signed-ranks test (Siegel, 1956), a nonparametric technique, was used to test the relative magnitude as well as the direction of the differences in item difficulty of the selected items for the students having encountered the item on a CAM test form prior to encountering it on a unit posttest and those students with no prior encounter with the item before the unit posttest. The null hypothesis was that there was no difference between the two

groups as to performance on the 10 items,

$$H_0 : a_1 = a_2,$$

where a equals a group's performance on all 10 items.

A second test of differences between the two groups (which are termed the Prior CAM Experience Group and the No Prior CAM Experience Group) was made using the Walsh test (Siegel, 1956). An assumption of this test is that the difference scores observed in the two selected samples are drawn from symmetrical populations, thus mean and median are assumed to be equal. The null hypothesis is that the average of the difference scores (u_o) is zero. Stated in standard form the null hypothesis is

$$H_o : u_o = 0,$$

and the alternative hypothesis,

$$H_1 : u_o \neq 0$$

for a two-tailed test. The Walsh test, also a nonparametric technique, is a more conservative test than the Wilcoxon (Siegel, 1956).

3.2.1.2 Sample. The sample consisted of seven of the eight student schedule groups. The three student schedule groups with prior experience with the test item on a CAM test form before instruction on the associated objective were

collectively called the Prior CAM Experience Group. The four student schedule groups with no prior experience with the item were called the No Prior CAM Experience Group.

3.2.1.3 Data. Item difficulty was calculated with single student schedule groups as the base. Ten items were selected randomly for the analysis from Unit 26 and appear on the CAM test forms and on Unit Test 63. (Appendix A contains a breakdown of both CAM test forms and unit posttest forms by items and by objectives to which the items are associated.) There were several reasons for selecting these items:

(1) Unit Test 63 follows four CAM test administrations allowing the test items of Unit 26 to be in a preinstruction phase for the first three CAM administrations.

(2) Unit Test 63 was administered on 17 December, 1971, ten days after instruction on Unit 26 had been completed (7 December, 1971) providing a fairly immediate posttesting.

(3) A CAM test administration occurred 8 December, 1971, which permitted posttesting of Unit 26 by CAM forms prior to Unit Test 63. Thereby both preinstruction and postinstruction data were provided for the items on the CAM forms which allowed comparison to the Unit Test 63 postin-

struction data.

(4) Unit 26 has eight objectives all of which are tested on Unit Test 63. The objectives are tested by a total of 24 items, 10 of which also appear on the CAM forms and which are the focus of this analysis.

(5) The eight objectives of Unit 26 are tested by several items each on Unit Test 63, unlike most other units which have only one item per objective on a unit posttest.

(6) The 10 items which appear on both CAM forms and Unit Test 63 were encountered for the first time by some students on Unit Test 63 because there were two CAM test administrations following the administration of Unit Test 63. This provided a group of students who did not encounter all of the 10 items after Unit Test 63. This was true because CAM Test Forms 56 and 57 did not contain any of the 10 items, and in addition, since only four CAM test administrations occurred before Unit Test 63 was administered, there were still four CAM test forms that had not been administered to each group. Thus, each student schedule group had encountered only from two to eight of the 10 items prior to Unit Test 63.

3.2.1.4 Results. Ten tables, each containing the item difficulties for one of the 10 items for each of the eight student schedule groups (abbreviated SSG in the tables), were constructed. For brevity only one of the 10 tables is presented here. All 10 tables are presented in Appendix B for reference. (In Table 3.2.1 below, the test form identification number is given directly below the test administration heading.)

Test Administrations 1, 2 and 4 occurred before instruction on the objective measured by Item 260101. Test Administrations 6, 7, 8 and 9 occurred after instruction. Item difficulty is given for Item 260101 on the given test form for the given test administration and for the particular student schedule group. The student schedule groups, it should be recalled, were unequal in size ranging from 29 to 36 students each. Table 3.2.1 is similar to the other nine tables of Appendix B. Two of the eight student schedule groups did not have a chance to encounter a given particular item on a CAM test form due to the particular sequence in which the CAM forms were administered and because each group only received six of eight of the test forms. Two other student schedule groups encountered each item on a CAM test form after Unit Test 63. The other four student schedule groups encountered the item on CAM tests prior to Unit Test 63.

Examining Tables 3.2.1 and B-2 through B-10 of Appendix B, it was found in six of the 10 tables that the three student schedule groups who encountered the item on a CAM

TABLE 3.2.1

Item Difficulty of Item 260101 by Student Schedule Group (SSG) and by Test Administration

SSG (and Size)	Test Administration 1 CAM 51	2 CAM 51	4 CAM 51	6 CAM 51	7 Unit 63	8 CAM 51	9 CAM 51
1 (29)	.21				.83		
2 (30)			.23		.83		
3 (36)					.86		
4 (33)				.56	.88		
5 (31)					.87	.71	
6 (35)		.31			.94		
7 (33)					.88		.79
8 (29)					.82		

Note.--The type of test and the test form identification number are given immediately below the test administration number.

test before instruction on the associated objective performed better on that item on Unit Test 63 than the four student schedule groups who encountered the item for the first time on Unit Test 63. However, the differences were generally slight. Table 3.2.2 displays the item difficulties for the 10 items for the two treatment groups.

The input for both the Wilcoxon match-pairs signed-ranks test and the Walsh test for differences of mean item

difficulty between the Prior and No Prior CAM Experience Groups is presented in Table 3.2.3. From the Wilcoxon test, the difference between the sum of the ranks prefixed by a plus sign and the sum of the ranks prefixed by a minus sign was "1" which was not significant at the .05 level of confidence and the null hypothesis was not rejected. This is evidence of no difference between the groups. From the Walsh test, the difference was also found to be nonsignificant at the .05 level of confidence and, therefore, the null hypothesis of no difference was also not rejected.

Interpreting the results of both tests, the conclusion is drawn that the effect of encountering one of the 10 selected items on a CAM test before instruction on the objective which the item measures generally has no affect on the students' ability to answer the same item when encountered again on Unit Test 63 after instruction on the associated objective. This finding supports the assertion that encountering items on a CAM test before instruction on the associated objective does not have an effect on performance when the items are encountered again in a postinstruction phase on a unit test. One of the claims of the CAM literature is that student exposure to testing of objectives before instruction occurs will encourage students to study these objectives before final instruction. The findings for Question 2A do not support this assertion.

TABLE 3.2.2

Item Difficulty of the 10 Items on Unit Test 63 for the Prior CAM Experience Group and for the No Prior CAM Experience Group

Item Number	Prior CAM Experience Group	No Prior CAM Experience Group
260101	.87 (N = 94)	.86 (N = 129)
260110	.78 (N = 97)	.88 (N = 129)
260204	.84 (N = 100)	.72 (N = 121)
260303	.72 (N = 97)	.79 (N = 129)
260304	.85 (N = 98)	.82 (N = 129)
260401	.65 (N = 93)	.74 (N = 134)
260402	.64 (N = 95)	.58 (N = 128)
260501	.75 (N = 94)	.77 (N = 131)
260702	.67 (N = 95)	.66 (N = 128)
260705	.38 (N = 94)	.34 (N = 129)

TABLE 3.2.3

Input for the Wilcoxon Matched-Pairs Signed-Ranks Test and the Walsh Test for Differences of Item Difficulty Between the Prior CAM Experience Group and the No Prior CAM Experience Group

Item Number	Difference Between Prior CAM Experience Group and No Prior CAM Experience Group (to Four Decimal Places)	For Wilcoxon Test: Rank of Difference by Absolute Size Prefixed by Sign of Difference (Smallest Difference = 1)		For Walsh Test: Rank in Order of Size (Smallest = 1)
260101	+.0130	+2		6
260110	-.1025	-9	Sum of ranks	1
260204	+.1234	+10	with plus:	10
260303	-.0720	-7	28	3
260304	+.0266	+4		7
260401	-.0855	-8	Sum of ranks	2
260402	+.0595	+6	with minus:	9
260501	-.0149	-3	27	4
260702	+.0122	+1		5
260705	+.0470	+5		8

3.2.2 Analysis of Question 2 (B)

Question 2B focused on whether or not item difficulty depends on the context in which the item is encountered. In this instance, the contexts were the CAM tests and the unit posttests.

3.2.2.1 Design. As with Question 2A both the Wilcoxon matched-pairs signed-ranks test and the Walsh test were employed to test for differences in performance between the two treatment groups. The groups were: (1) those students encountering the 10 items first on a CAM test in a postinstruction phase (CAM Postinstruction Experience Group) and (2) those students encountering the items for the first time on Unit Test 63 (No Prior CAM Experience Group).

3.2.2.2 Sample. The sample consisted of the single student schedule group which encountered the 10 items for the first time on a CAM test in a postinstruction phase and the four student schedule groups which first encountered the items on Unit Test 63.

3.2.2.3 Data. As with Question 2A, the data were item difficulties by student schedule group and test form.

3.2.2.4 Results. Table 3.2.4 summarizes the input data for both the Wilcoxon and Walsh tests. Examining the differences, it is seen that eight of 10 differences are negative. This means that performance on eight of 10 items

TABLE 3.2.4

Difference Scores for the 10 Items Between the CAM Postinstruction Experience Group and the No Prior CAM Experience Group and the Rankings for the Wilcoxon and Walsh Tests

Item Number	Difference Between CAM Postinstruction Experience Group and No Prior CAM Experience Group (to Four Decimal Places)	For Wilcoxon Test: Rank of Difference by Absolute Size Prefixed by Sign of Difference (Smallest Difference = 1)		For Walsh Test: Rank in Order of Size (Smallest = 1)
26101	-.3005	-10		1
260110	-.1740	-7	Sum of ranks	4
260204	+.1120	+4	with plus:	10
260303	-.1778	-8	5	3
260304	-.0617	-3		7
260401	-.2235	-9	Sum of ranks	2
260403	-.1584	-6	with minus:	5
260501	-.1546	-5	50	6
260702	+.0053	+1		9
260705	-.0124	-2		8

was poorer on CAM tests in a postinstruction phase than on Unit Test 63. For both the Wilcoxon matched-pairs signed-ranks test and Walsh test, the null hypothesis of equal means was rejected at the .02 level of confidence. Clearly, the item difficulty implies that the students performed more poorly on the same item when it was encountered after instruction on a CAM test than on the unit posttest. This conclusion is further reinforced by examining the 10 tables of Appendix B. For six of the 10 items, the difficulty increased for the CAM postinstruction Test Administrations 8 and 10 over the unit posttest item difficulties of Test Administration 7. For two of the other items the CAM postinstruction item difficulty decreased very slightly from the unit posttest item difficulty.

3.2.3 Discussion

Analysis of Question 2A showed that students who have encountered an item prior to instruction on the associated objective on a CAM test score the same on the item on a unit posttest as students who have not encountered the item on a CAM test prior to instruction on the associated objective. Several possible reasons for this performance follow.

First-year high school biology contains much material new to students. Most students have limited preknowledge. The specific objectives of the course used in this study were generally tied to textbook material in the course. Students would have had to read ahead and remember content of the text-

book to benefit from encountering a test item prior to instruction. Items encountered prior to instruction would have been new and not likely to have been remembered easily. Other objectives were dependent on laboratory experiences which students would not have had previously.

It appears that the context of the CAM test in which one encounters only a small number of items within an immediate postinstruction phase and a large number of preinstruction or long-term postinstruction items, depending on when the test is administered, tends to limit the effects of associating a pretest item with other items on a test which may have familiar context. A pretest item would tend to be less closely related to other items on the test form because of the more course-representative nature of the test compared to a unit posttest. The student would then be at a disadvantage compared to a unit posttest situation in deriving clues from some items to answer others. Further study with other subject areas and student age levels are needed for conclusive evidence of the effects of encountering an item twice in two different contexts.

The analysis of Question 2B examined the effects of the two testing contexts, CAM testing and unit posttesting. The conclusion of significant differences between performance in the CAM and unit test contexts with students doing more poorly on the items when encountered on CAM tests must be qualified. In addition to differences which may be attributed to different test contexts are differences due to moti-

vation may be the most important cause of difference. For the unit posttest, students have the time and motivation to study a defined segment of material for several days after instruction. For the CAM test, the defined material spans the semester so the motivation to do well on the CAM test would tend to be less than for a unit test. This may be a shortcoming of the CAM longitudinal design. It may also be a problem due to the amount of importance the teacher places on the CAM tests for purposes of assigning grades and using the information for program refinement versus the importance placed on the unit tests. It is evident that unit tests are more important to both the teacher and students.

The factor of test context should be studied further. More studies with other course subject areas and student populations need to be conducted for conclusive evidence, but for this course it would appear that the problem of identifying an item on a CAM test as one which should be known from instruction as opposed to other kinds of items (pretest and perhaps long-term retention items) is a factor in answering the item correctly. The unit posttest context of solely postinstruction items relating to one or two units of study may have a significant effect on one's item performance as contrasted with a context of less closely related items. It may be true that verbal cues are more common with items having related content and that being able to answer certain items may also lead to solutions of other items related to the few objectives being measured in a unit test context.

Thus, CAM-type tests, that is criterion-referenced tests administered longitudinally with course-representative content, furnish conservative estimates of achievement in comparison with criterion-referenced tests which are more posttest oriented and contain items related to fewer objectives.

3.3 Question (3)

> What are the correlates of various ability measures with several global measures of the success of individual students in a course?

The relationships among global measures of course success and predictors of success are useful indicators of which likely predictors actually could be used as such. Likely predictors which correlate highly with each other are measuring a similar thing and therefore the combined use of them in prediction of a global measure of success will not improve the prediction appreciably. Those which do not correlate highly with each other may improve the prediction of a global measure.

3.3.1 Analysis of Question 3

3.3.1.1 Design. The analysis was a series of correlations among likely predictors and global measures of course success (i.e., the criterion measures for prediction). The results are provided in an intercorrelation matrix of the six selected global measure variables and the eight selected likely predictor variables. The six global measure

variables are:

(1) sum of the unit posttest scores,

(2) sum of CAM total test scores,

(3) sum of CAM preinstruction scores (not really a measure of course success, but this was included for comparison with other measures),

(4) sum of CAM instruction-completed scores,

(5) teacher's final grade for semester, and

(6) CAM test score of the final CAM test administration.

The eight predictor variables are as follows:

(1) SCAT II verbal subtest score (available for 208 of the 256 students),

(2) SCAT presence/absence score (a "0" if the student did not have a SCAT score and a "1" if he did),

(3) normalized SCAT verbal subtest score determined for each subject by subtracting the group mean of the SCAT score from each subject's SCAT score,

(4) first CAM test score,

(5) second CAM test score,

(6) first unit posttest score,

(7) sum of first two CAM test scores, and

(8) sum of first three test scores; i.e., first two CAM test scores and the first unit posttest score.

3.3.1.2 Sample and data used. The test scores of all 256 students were used for the correlational analyses with the exception of the raw SCAT II verbal test scores which were from the 208 students who had them.

3.3.1.3 Results. Table 3.3.1 contains an intercorrelation matrix of the 14 variables. The SCAT score was included as a possible predictor variable because SCAT scores: (1) were known by teachers before either the CAM or unit test scores and (2) are a measure of general varbal ability. The dichotomous variable "presence/absence of SCAT score" was included to determine whether students with SCAT scores differed from those with SCAT scores. A slight difference in the two groups is shown by the consistent trend of the correlations with SCAT P/A to be small (.00 to .09) but positive.

The normalized SCAT score was determined by subtracting the group mean of the SCAT score, which was 73.244, from each subject's SCAT score. For those students who had no SCAT scores the group mean was assigned. Thus, when the group mean was subtracted from each score, the normalized scores for these students were zeroes. The normalized SCAT scores produced correlations with the global measures of success which were only slightly lower than the correlations produced with the raw SCAT scores.

A derived score (the normalized SCAT score) was used so that all 256 students would have a SCAT score. Therefore, only a slight difference between students with SCAT scores and those without the normalized SCAT scores can be used in place of the raw SCAT score in prediction equations. Since the correlation produced with the normalized scores were lower than those produced with the raw scores, they may be of more conservative power for prediction than the raw SCAT scores.

TABLE 3.3.1

Intercorrelation Matrix of Course Success Predictor Variables and Course Global Measures of Success[c]

Variables		Variable Identification Number												
ID Number	Name	1	2	3	4	5	6	7	8	9	10	11	12	13
1.	SCAT Score													
2.	SCAT P/A	.00												
3.	Normalized SCAT	1.00	.00											
4.	1st CAM Score	.41	.08	.36										
5.	2nd CAM Score	.29	.05	.27	.18									
6.	1st Unit Score	.59	.04	.54	.35	.29								
7.	1st and 2nd CAM	.45	.08	.41	.72	.81	.41							
8.	1 CAM, 2 CAM, 1 UNIT	.61	.07	.57	.64	.65	.85	.84						
9.	Sum of Unit[a]	.65	.09	.59	.42	.30	.83	.47	.77					
10.	Sum of CAM[a]	.66	.08	.59	.55	.55	.68	.72	.83	.83				
11.	Sum CAM Preinst.[a]	.45	.05	.40	.55	.76	.37	.87	.73	.44	.77			
12.	Sum CAM Postinst.[a]	.64	.08	.58	.43	.30	.72	.47	.71	.87	.92	.45		
13.	Final Grade[a,b]	.45	.09	.49	.36	.30	.69	.42	.66	.85	.74	.40	.78	
14.	Final CAM Score[a]	.59	.04	.53	.37	.25	.64	.39	.61	.78	.80	.35	.90	.70

[a]Global measures of success.

[b]Final grade was recorded numerically as: A = 5, B = 4, C = 3, D = 2, F = 1.

[c]Raw SCAT score correlation based on N of 208, all others based on N of 256.

Examining each of the other possible predictor variables, the first CAM test score had its highest correlation with the sum of the CAM test scores and the sum of the CAM preinstruction scores, both being .55. The second CAM score had its two highest correlations with the same two global measures, .55 and .76 respectively, as should be expected due to the redundancy in measurement among the CAM tests. The first unit test score correlated highly with five of the six global measures with coefficients ranging from .64 to .83 for the five measures. The low coefficient of .37 was for the sum of CAM preinstruction scores and would be expected to be lower than the others because all the other variables were measures of course success and were composed largely or completely of postinstruction data.

It should be pointed out that many of the high correlations are those of global measures with individual test scores which were in fact components of the global measures.

The sum of first and second CAM test scores had higher correlations with all six global measures than either test score separately. This was expected since the sum is more reliable than individual scores. However, the first unit test score correlated more highly with four of the six global measures (the sum of the CAM test scores and the CAM preinstruction scores were the two exceptions) than did the sum of the first and second CAM test scores. The addition of the first unit test score to the two CAM test scores further increased the correlations an average of .22 for five of the

six global measures. The exception was the sum of the CAM preinstruction scores where a drop from .87 to .73 occurred from the correlation with the sum of the first two CAM test scores to the sum of the first two CAM test scores plus the first unit test score.

It is apparent that the first unit test score alone and the raw SCAT score (or normalized SCAT score for N of 256) are likely to be the best of the available noncomposite predictors of the six global measures. The combination of first and second CAM test scores and the first unit test score may prove the best predictor of the eight offered and perhaps a very good predictor in its own right. The measures most indicative of course success (i.e., sum of CAM instruction-completed scores, sum of unit test scores, and, of course, final course grade) correlate .71, .83, and .66 respectively with the combined sum of the two CAM and one unit test scores. Such high correlations are likely to be indicative of good predictors of these global success variables.

The most striking information from Table 3.3.1 is summarized below:

(1) There is an absence of negative correlations, with most correlations being above .30 (with the expected exceptions of the correlations of the SCAT P/A variable).

(2) Generally, very high correlations (most between .61 and .83) were produced with the sum of the

three test scores or with the unit test score alone. This was undoubtedly due to the situation of predictor variable and criterion variable (i.e., measure of global success) each containing an overlap of information since the items of the predictor were often shared with the final criterion measure.

(3) The correlation between either the first CAM test score, second CAM test score, or the sum of the two CAM test scores and the global measures were generally lower than those produced with the unit test score alone or added to the two CAM scores.

3.3.2 Discussion

The rationale behind Question 3 was to determine if CAM test data furnish information valuable apart from unit test data in regard to prediction of final success in the course. Question 3 identified the possible good predictor variables and Question 4 examined their value in predicting the global measures.

A basic problem to the comparison of the global measures and test scores is the overlap among these variables. Since test items of both predictor and criterion variables were shared in many instances, the correlations produced were contaminated to an unknown degree. Therefore, no clear con-

clusions can be drawn for Question 3. Other factors may enter into the production of differences and these are more fully examined below.

Several factors in addition to contamination from shared items may partially explain the differences between the magnitude of correlations between most of the global measures and the CAM test scores which were generally much less than the correlations between the global measures and the unit test scores. First, the biology course is one requiring a specialized knowledge, little of which is likely to have been presented in previous school courses allowing few students to have much preknowledge and probably few who push ahead of instruction on their own. Second, the course is based on reading a text and on laboratory experiences. Thus, students who do better on unit tests will be the better readers and will have better study habits. Also the laboratory experiences do not allow students to work ahead. Third, the stated behavioral objectives for the course are largely based on textbook readings. This would also be part of the explanation for the high correlation between SCAT scores and course success measures.

Another problem is the difference in test length between CAM and the unit tests. If the CAM postinstruction correlations were corrected for unreliability (i.e., attenuation), the resulting comparisons might be different from those obtained.

Final grades also presented a problem since they were

determined by each of the four teachers for their own students. They were based on CAM and unit test results (two-thirds) and homework and laboratory performance (one-third). Thus, final course success as reflected by final semester grade was to a large degree directly based on performance on tests and work (both laboratory and homework) which was dependent on and matched the instructional pace set for the group by the teacher. Therefore, better readers and students geared to the pace set by the teacher would be at an advantage in answering items related to objectives encountered in a postinstruction phase.

Finally, it should be pointed out that the correlations reported in Table 3.3.1 were computed using the total N of 256 (with the exception of the raw SCAT scores where N was 208). If the two teacher sections where CAM was irregularly used were not included (reducing N to 217), the correlations with the second CAM administration score in particular might be higher (see analysis of Question 1).

3.4 Question (4)

Do early CAM test scores predict global measures of individual student progress and final semester grade?

The analysis of this question deals with the variables used in the correlation matrix in Table 3.3.3 of the analysis of Question 3. The focus is twofold: (1) to determine if early CAM test scores predict global measures of

individual student progress and final semester grade and (2) to compare the value of CAM test scores as predictors with and without consideration of other predictor variables such as the first unit test score and the SCAT II verbal score. The second focus includes the trial use of CAM test scores and other predictor variables in simulated decision situations in order to test the practicality of using CAM test scores in prediction in actual classroom settings.

3.4.1 Analysis for Focus (1)

The analyses conducted were multiple-regression analyses wherein a linear combination of independent variables is produced which correlates as highly as possible with the dependent variables.

3.4.1.1 Design. The general multiple regression prediction equation was:

$$\hat{Y}_i = b_o + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_n X_{in} \qquad (4 - 1)$$

where, $\hat{Y}_i$ = the predicted dependent variable for subject i,

b_o = an additive constant (point where the linear regression plane intercepts the vertical axis),

b_j = the regression coefficient of the jth independent variable X_j (where j = 1, ..., n), and

x_{ij} = the score on predictor (i.e., independent) variable j of subject i.

Stepwise multiple regression is the variation of multiple regression used for the analyses of Question 4. In

ordinary stepwise multiple regression, independent variables are selected from those available to construct at each step that regression equation which provides the best prediction possible with a given number of independent variables. Each new equation containing another variable not included in previous equations is termed a "step." Each step builds on the previous step in forming a series of equations, each containing one more variable than the previous until no variables are left which will increase the multiple correlation coefficient, R, significantly.

Variables may also be "forced" into equations in any desired order. By this method a given variable is placed in an equation by the user of the computer program. "Forcing" enables estimation of the relative value of a variable in an equation containing a particular "context" of other variables which might not naturally occur if the variable were "free" to enter or not "free" to enter at any step.

The null hypothesis of interest for the analysis of Question 4 was that either the regression coefficent is zero or that the increase in the multiple R is zero. In proper notation the hypothesis was:

$$H_o : b_{ij} = 0 \text{ or } R_{ij} = 0$$

where, b_{ij} = the regression coefficient of the ith variable included in the equation for independent variable j, and

R_{ij} = the increase in the multiple R due to the inclusion of the ith variable in the equa-

tion for independent variable j over the R of the equation containing variables 1, 2, ..., i-1.

The test of this hypothesis is an F-test, the results of which are reported for each inclusion of a variable and then for its subsequent possible deletion.

The intercorrelation matrix of Question 3 provided a picture of relationships among the independent variables themselves and among the independent and dependent variables which provided aids in selecting variables to be forces into equations and sequences in which to force them. The BMD02R computer program for stepwise regression was used to generate the prediction equations (UCLA, 1964d).

The most valid indicator of course success was probably the semester final grade. However, it should be stressed that this grade was constructed as being two-thirds dependent on CAM and unit test scores. Thus, there is no true independent indicator of success in the course. For the variable "semester final grade," prediction equations were generated in some detail. The following six variables were employed as predictors (i.e., independent variables) as suggested in Question 3:

(1) first CAM test score (1ST CAM);

(2) second CAM test score (2ND CAM);

(3) first unit test score (1ST UNIT);

(4) sum of the first and second CAM test scores (1+2 CAM);

(5) sum of the first CAM test, second CAM test, and first unit test scores (1C2C1U); and

(6) normalized SCAT score (SCATNM).

The capitalized terms in parentheses are the abbreviations for the variables used throughout the remainder of the report of the analyses for Question 4.

3.4.1.2 Sample and data used. The test scores of all 256 students were used. The normalized SCAT score was used in place of the raw SCAT score to include the 48 students without SCAT scores.

3.4.1.3 Results. In the tables which follow for each of the series of stepwise regressions reported, several column headings need expanded explanations. The column headed "F" designates one of two kinds of F-ratio. The F-ratio for the first variable entered into the equation is from the standard F-test. It is calculated before entry of the variable. The following F-ratios reported in the column for each successive variable included in the equation are from the sequential F-test which makes allowance for variables already entered into the equation. It is also calculated before entry of the variable into the equation and serves the function of determining whether the variable in question should be entered into the prediction equation. Other F-values are also reported in the tables. The numbers in parentheses under the "F" column are F-ratios from the partial F-tests for the elimination of variables already in

the regression equation. For this particular study the minimum F-level for entry of a variable was the .01 level of confidence and for elimination .005, both values being the default values for the BMDO2R program.

As an example of the analysis of variance summary provided for each calculation of an F-ratio for entry of a variable into the equation, the summary table for the analysis of variance for the entry of the first variable into the first prediction equation of the stepwise regression summarized in Table 3.4.1.2 is shown below in Table 3.4.1.1.

For ease in reading the tables below which summarize the results of the stepwise regressions for the several dependent variables, the following explanation of table labelling is presented. The column headed "Step number" designates the steps in sequence of the particular stepwise regression. "Independent variable(s)" refers to the variable or variables in the equation. The number in parentheses following the variable name is the position of the variable in the equation (i.e., position entered into the equation). "F" has been explained previously. "Multiple R" is the multiple correlation coefficient of the dependent variable with the independent variables of the equation. "R^2" is the square of the multiple R and is the proportion of variance in common between the dependent variable and the independent variables of the equation. Thus, it is the shared variance of Y and X_1, ..., X_n with "n" designating the number of independent variables. The differences between the R^2 values

TABLE 3.4.1.1

Summary of the Analysis of Variance for the Entry of the Variable 1ST UNIT as the First Free-to-Enter Variable in the Prediction Equation for the Dependent Variable of Semester Final Grade

Source of Variation	df	Sum of Squares	Mean Square	F-ratio
Total	255	359.24		
Regression (1ST UNIT)	1	170.15	170.15	228.55
Residual	254	189.09	.74	

TABLE 3.4.1.2

Results of Stepwise Regression for the Dependent Variable of Semester Final Grade with Independent Variables Free to Enter

Step Number	Independent Variable(s)	F		Multiple R	R^2	b_o	b_1	b_2	b_3
1	1ST UNIT	228.55		.69	.47	- .75	.14		
2	1ST UNIT(1)		(159.51)						
	1+2 CAM(2)	124.96	(11.72)	.70	.50	-1.25	.12	.04	
3	1ST UNIT(1)		(107.50)						
	1+2 CAM(2)		(7.70)						
	SCATNM(3)	86.69	(5.61)	.71	.51	- .75	.11	.03	.01

in the column are a measure of the relative value for prediction purposes of the variables added in the respective positions to the equation (Darlington, 1968). The column head "b_o" is the constant of the equation and the columns headed "b_1," "b_2," ..., "b_n" designate the regression coefficients of the respective variables as entered into the equation.

Table 3.4.1.2 summarizes the stepwise regression results for the dependent variable of semester final grade with the six independent variables, all of which are free to enter the prediction equation.

Examination of Table 3.4.1.2 shows that the first unit test score when entered first accounts for 47% of the shared variance and the addition of the next two variables only increases R^2 to .51. Thus, once the first unit test score is available it overshadows all other variables as the single best predictor; other variables added to it increase prediction accuracy very little.

Since the teachers in the study had available the SCAT scores and the first and second CAM test scores before the first unit test score, the four variables were forced in order of their chronological availability. The series of equations formed might be of the type possibly useful to a teacher in a similar situation to make predictions. Table 3.4.1.3 summarizes the stepwise regression results for the four variables forced in sequence of availability to teacher. The importance of the first unit test score in predicting

TABLE 3.4.1.3

Results of Stepwise Regression Analysis for the Dependent Variable of Semester Final Grade and Four Forced Independent Variables

Step Number	Independent Variable(s)	F		Multiple R	R^2	b_o	b_1	b_2	b_3	b_4
1	SCATNM	79.09		.49	.24	3.44	.03	-	-	-
2	1ST CAM(1)	48.46	(13.84)							
	SCATNM(2)		(51.21)	.53	.28	2.50	.07	.02	-	-
3	1ST CAM(1)		(12.37)							
	2ND CAM(2)		(8.23)							
	SCATNM(3)	35.98	(40.99)	.55	.30	1.91	.07	.05	.02	-
4	1ST CAM(1)		(4.78)							
	2ND CAM(2)		(2.84)							
	1ST UNIT(3)	64.94	(106.60)							
	SCATNM(4)		(5.36)	.71	.51	-.76	.04	.02	.11	.01

the final semester grade is clearly shown in the table. The first three variables entered in chronological order of availability account for 30% of the shared variance, while the addition of the first unit test score increases the amount of variance accounted for to 51%.

Table 3.4.1.4 summarizes the stepwise regression results for the first three tests forced in sequence but without SCAT scores considered. Again, it can be seen that the first unit test score is the largest factor in the prediction of final semester grade. Given the unit test score, the value of the normalized SCAT score is almost completely removed (R^2 of .51 with SCATNM and .50 without). R^2 jumped from .19 for the two CAM test scores to .50 with the unit test score included. The F for deletion of the first unit test score in step 3 of Table 3.4.1.4 reflects the dramatic effect of adding the first unit test score to the prediction equation.

A major focus of this study was to investigate the relationship of CAM and unit testing, particularly in regard to the possibility of overlap in information provided. One approach is to determine the value of CAM in predicting unit test results. Therefore, stepwise regression analyses conducted with the dependent variable of the sum of unit test scores and the independent variables of the first and second CAM test scores furnish information relevant to the investigation.

When the six independent variables were free to enter

TABLE 3.4.1.4

Results of Stepwise Regression Analysis for the Dependent Variable of Semester Final Grade and the Variables of First CAM Test Score, Second CAM Test Score, and First Unit Test Score

Test Number	Independent Variable(s)	F	Multiple R	R^2	b_o	b_1	b_2	b_3
1	1ST CAM	38.17	.36	.13	1.85	.13	-	-
2	1ST CAM(1)	(30.79)						
	2ND CAM(2)	28.90 (17.20)	.43	.19	1.05	.11	.07	-
3	1ST CAM(1)	(7.34)						
	2ND CAM(2)	(3.97)						
	1ST UNIT(3)	83.36 (156.70)	.71	.50	-1.26	.03	.12	

the equation, the best single predictor was, as expected, the first unit test score, which correlates .83 with the sum of the unit test scores. However, when the variables are forced into the equation in chronological order of availability, the results are as reported in Table 3.4.1.5.

After the entry of SCATNM, the increase in R^2 was slight for each of the first two CAM test scores. As expected, the entry of 1ST UNIT dramatically increased R^2, the proportion of variance shared by the dependent variable and the combination of four independent variables.

Table 3.4.1.6 summarizes the results of stepwise regression when SCATNM is not entered until after the other three variables are taken in order of availability.

It can be seen from Tables 3.4.1.5 and 3.4.1.6 that

TABLE 3.4.1.5

Results of Stepwise Regression with Dependent Variable of Sum of Unit Test Scores and Independent Variables Forced in Order of Availability

Number of Variables in Equation	Variables in Order of Entry into Equation	Multiple R	R^2
1	SCATNM	.59	.35
2	1ST CAM	.63	.40
3	2ND CAM	.64	.42
4	1ST UNIT	.85	.73

TABLE 3.4.1.6

Results of Stepwise Regression Analysis with the Dependent Variable of Sum of Unit Test Scores and Three Forced Independent Variables

Number of Variables in Equation	Variables in Order of Entry into Equation	Multiple R	R^2
1	1ST CAM	.42	.18
2	2ND CAM	.48	.23
3	1ST UNIT	.84	.71
4	SCATNM	.85	.73

the first and second CAM test scores are weak predictors of the sum of the unit test score as compared to the SCATNM or the 1ST UNIT variable.

One further set of regression analyses is directly related to the study of the relation between CAM testing and unit testing. This is the set having the sum of CAM instruction-completed (i.e., postinstruction) scores as the dependent variable. This variable is the CAM counterpart to unit testing. Table 3.4.1.7 is a summary of several stepwise regression analyses, each of which had the six independent variables in a different combination of entry priorities.

Results are very similar to those of the stepwise regressions with the dependent variable of sum of unit test scores. The same four independent variables were present and multiple R's were highly similar with the exception of those of 1ST UNIT in each equation, where with the dependent variable of sum of unit test scores it ranged from .83 to .85 and in Table 3.4.1.7 it ranged from .72 to .77. Not only was 1ST UNIT the single best predictor but it eclipsed the other variables so that even with the others added into the equations R only increased from .72 to .77.

3.4.1.4 Summary of results. The "usefulness" of a predictor variable, as defined by Darlington (1968), is the amount that R^2 decreases if that predictor variable were removed from the regression equation and the remaining variables reweighted appropriately. When predictor variables are

TABLE 3.4.1.7

Results of Stepwise Regression Analyses with the Dependent Variable of Sum of CAM Instruction-Completed Scores and Several Combinations of Forced and Free-to-Enter Independent Variables

Restrictions on the Variables to Enter Equations	Number of Variables in Equation	Variables in Order of Entry into Equation	Multiple R	R^2
All free to enter	1	1ST UNIT	.72	.52
	2	SCATNM	.75	.57
	3	1ST CAM	.77	.59
	4	2ND CAM	.77	.59
All forced in order of availability	1	SCATNM	.58	.33
	2	1ST CAM	.62	.39
	3	2ND CAM	.64	.40
	4	1ST UNIT	.77	.59
Same as above but SCATNM left free to enter after 1, 2 and 3 were in	1	1ST CAM	.43	.18
	2	2ND CAM	.49	.24
	3	1ST UNIT	.75	.56
	4	SCATNM	.77	.59

uncorrelated, then the usefulness of a given predictor variable equals the squared correlation of the dependent variable and the predictor variables. When predictor variables are intercorrelated, the usefulness of a given predictor variable equals the squared correlation of the dependent variable and that component of the predictor variable which is orthogonal to the other predictor variables.

The correlation matrix in Table 3.3.3 of the analyses results of Question 3 show that all the predictor variables of this study are intercorrelated and thus it is the orthogonal component of each which is used to determine usefulness in Darlington's sense.

In determining the usefulness of predictor variables reported in this analysis, it was found that the first unit test score (1ST UNIT) was the most useful in predicting course semester grade in every equation. The 1ST UNIT alone provided an R^2 of .47 (see Table 3.4.1.2) and the normalized SCAT score (SCATNM) alone provided an R^2 of .24. The first CAM test score (1ST CAM) alone provided an R^2 of only .19. However, adding 1STUNIT provided an R^2 of .47 and adding other independent variables increased R^2 to only .51. In contrast, the three variables of SCATNM, 1STCAM, and 2NDCAM provided an equation with R^2 of .30.

Thus, the analysis showed that early CAM scores in this study were poor predictors of final semester grade while normalized SCAT score and first unit test scores were more useful in the prediction equation.

3.4.2 Analysis for Focus (2)

A second part of the analysis for Question 4 was the usefulness of several of the prediction equations to the teacher making decisions about each individual student's course progress. The most indicative variable (of the available variables) of overall course success was the final semester grade. Thus this variable was used as the dependent variable in all equations for this part of the analysis.

3.4.2.1 Design. Simulated decision making was conducted and decision "boxes" were constructed for each combination of student group and prediction equation selected. Decisions were based on the actual final semester grade versus that predicted by the particular equation. Each box contains four cells. Each cell contains the sum of decision outcomes of a particular type which is one of the following four:

	Prediction	Actual Result
(1)	Pass	Pass
(2)	Pass	Fail
(3)	Fail	Pass
(4)	Fail	Fail

The "proportion of correct decisions," i.e., types (1) and (4), was calculated for each prediction as well as for decision types (2) and (3). The loss function selected for comparison of the decision rules was the "proportion of incorrect decisions."

Prediction equations were calculated for the entire group and for all students with even-numbered student identification numbers. The odd-even division produced two random population halves. (Student identification numbers are six-digit numbers assigned to students upon entry into the schools of the district of this study.) The equations based on the even-numbered students were used to make predictions about the odd-numbered students. Thus, the comparison of the "evens" and "odds" provided cross-validation of the usefulness of the prediction equations. The BMD03R multiple-regression computer program was used to generate the prediction equations (UCLA, 1964b). For decision-making purposes, a failing grade for a student was set at 2.5 or less and a passing grade at greater than 2.5. These grade points were determined by assigning a weight of 5 to an "A," 4 to a "B," 3 to a "C," 2 to a "D," and 1 to an "F." The point of 2.5 was selected as the cutoff because of the nearness of a grade below "C" to failure.

Each set of equations take the form of:

$$\hat{Y}_i = b_o + b_1 X_{ij} + b_2 X_{i2} + \ldots + b_j X_{ij}, \qquad (4 - 2)$$

where $\hat{Y}_i$ = the predicted grade for individual i,

b_o = the equation constant term,

b_j = the regression coefficient for the independent variable j, and

X_{ij} = the score of individual i on the independent variable j.

3.4.2.2 Sample. This included three student groups: all students (N = 256), students with even identification numbers (N = 128), and students with odd identification numbers (N = 128). It was coincidental that exactly half the students had even identification numbers and half had odd.

3.4.2.3 Results. The first set of equations contained the single independent variable of normalized SCAT score (SCATNM). For N = 256 the prediction equation was:

$$\hat{Y} = 3.440 + .029 \text{ (SCATNM)} \qquad (4 - 3)$$

Table 3.4.2.1 contains the decision box.

The percent in each cell designates what percent the number of decisions reported in the cell are of the type of predicted grade for that cell. Thus, in Table 3.4.2.1 for the N of 256, 233 students had a predicted grade of PASS but only 192 of them (82% of those predicted as passing) actually had a passing grade, while 41 of them (18%) actually had a failing grade. This compares to the base rate for the N of 256 of 79% who actually passed and 21% who actually failed. The percents for the prediction categories serve as a means of comparison for predictive value of a given equation.

The prediction equation for N of 128 (even-numbered students) was:

$$\hat{Y} = 3.435 + .029 \text{ (SCATNM)}. \qquad (4 - 4)$$

Table 3.4.2.1 contains the decision box matrices for all three

TABLE 3.4.2.1

Decisions for Prediction Equations with SCATNM as Independent Variable for N = 256, N = 128 (Even-Numbered), and N = 128 (Odd-Numbered)

Predicted Grade (N)	Actual Grade: Pass		Actual Grade: Fail	
	N	%	N	%
For N = 256				
Pass (233)	192	(82%)	41	(18%)
Fail (23)	9	(39%)	14	(61%)
	201	(79%)	55	(21%)
For N = 128 (even-numbered)				
Pass (122)	102	(84%)	20	(16%)
Fail (6)	3	(50%)	3	(50%)
	105	(82%)	23	(18%)
For N = 128 (odd-numbered)				
Pass (111)	90	(81%)	21	(19%)
Fail (17)	6	(35%)	11	(65%)
	96	(75%)	32	(25%)

student samples: the N of 256, the N of 128 of even-numbered students, and the N of 128 of odd-numbered students for which decisions were based on the equation derived for the even-numbered students.

Given the student's SCAT score and the prediction that the student will pass, the accuracy was 192 of 201 (96%) and the prediction that the student will fail, the accuracy was 14 of 55 (25%) for the N of 256. However, the 192 was only 82% of those predicted to pass and the 14 was 39% of those predicted to fail. Thus, the total of each group for the prediction categories of pass or fail was not a strongly reliable indication of the actual proportion of those who do pass or fail under the criteria set forth. The use of the even-numbered students (N = 128) changed the prediction capabilities very slightly from those based on N of 256 and cross-validation with the odd-numbered students (N = 128) substantiated the value of the prediction equation.

A second set of equations contained the three independent variables of SCATNM, 1STCAM, and 2NDCAM. The equation for N of 256 was:

$$\hat{Y} = 1.909 + .069\ (1STCAM) + .047\ (2NDCAM) + .022\ (SCATNM). \qquad (4 - 5)$$

The equation for N of 128 even-numbered students was:

$$\hat{Y} = 1.869 + .098\ (1STCAM) + .021\ (SCATNM). \qquad (4 - 6)$$

Table 3.4.2.2 contains the decision results for the three

groups of students.

Again, the subsample of even-numbered students provided similar predictions to those of the entire sample and the use of the odd-numbered students as a cross-validation of the even-numbered students also provided highly similar results. The results are almost identical to those of Table 3.4.2.1 where only SCATNM was used in the equations. Thus, the inclusion of early CAM scores with SCAT scores for prediction did not improve prediction capabilities over using the SCAT score alone.

A third set of prediction equations had four independent variables: SCATNM, 1STCAM, 2NDCAM, and 1STUNIT. The equation for N of 256 was:

$$\hat{Y} = -.763 + .037\ (1STCAM) + .024\ (2NDCAM) + .113\ (1STUNIT) + .008\ (SCATNM), \qquad (4 - 7)$$

and for N of 128 even-numbered students was:

$$\hat{Y} = -1.630 + .053\ (1STCAM) + 0\ (2NDCAM) + .145\ (1STUNIT) + .003\ (SCATNM). \qquad (4 - 8)$$

For computation of decision values SCATNM was omitted because of nonsignificance of its addition to the equation, i.e., its regression coefficient did not differ significantly from zero. The decision results for the three students groups are contained in Table 3.4.2.3.

Once again the results for the 128 even-numbered students are similar to those for the N of 256 and the comparison

TABLE 3.4.2.2

Decisions for Prediction Equations with SCATNM, 1STCAM, and 2NDCAM as the Independent Variables for N = 256, N = 128 (Even-Numbered), and N = 128 (Odd-Numbered)

Predicted Grade (N)	Actual Grade: Pass N	Pass %	Fail N	Fail %
For N = 256				
Pass (232)	192	(83%)	40	(17%)
Fail (24)	9	(37%)	15	(63%)
	201	(79%)	55	(21%)
For N + 128 (even-numbered)				
Pass (123)	102	(83%)	21	(17%)
Fail (5)	3	(60%)	2	(40%)
	105	(82%)	23	(18%)
For N = 128 (odd-numbered)				
Pass (109)	90	(83%)	19	(17%)
Fail (19)	6	(32%)	13	(68%)
	96	(75%)	32	(25%)

TABLE 3.4.2.3

Decisions for Prediction Equations with Four Independent Variables for N = 256, N = 128 (Even-Numbered), and N = 128 (Odd-Numbered)

Predicted Grade (N)	Actual Grade Pass		Fail	
	N	%	N	%
For N = 256				
Pass (221)	192	(87%)	29	(13%)
Fail (35)	9	(26%)	26	(74%)
	201	(79%)	55	(21%)
For N = 128 (even-numbered)				
Pass (110)	100	(91%)	10	(9%)
Fail (18)	5	(28%)	13	(72%)
	105	(82%)	23	(18%)
For N = 128 (odd-numbered)				
Pass (99)	85	(86%)	14	(14%)
Fail (29)	11	(38%)	18	(62%)
	96	(75%)	32	(25%)

for cross-validation with the 128 odd-numbered students using the equation based on the 128 even-numbered showed quite similar proportions with some slight shifts in the "predicted fail" category.

A final set of equations considered only the single independent variable of 1STUNIT. The equation for N - 256 was:

$$\hat{Y} = -.746 + .139\ (1STUNIT), \qquad (4 - 9)$$

and for the N = 128 even-numbered students was:

$$\hat{Y} = -1.450 + .161\ (1STUNIT). \qquad (4 - 10)$$

Table 3.4.2.4 contains the decision results for these two groups and the comparison sample of N = 128 odd-numbered students based on Equation 4 - 10.

Results are highly similar to those of the previous three sets of prediction equations with proportions of correct decisions being highly similar across all three groups.

3.4.2.4 Summary of results. Clearly emerging from these analyses is that the SCAT score known at the beginning of the course was (with prediction accuracy of 96%) as useful as the first unit test score in predicting success or as any combination of the other independent variables. However, it was not as useful in predicting failures as the first unit test or other combinations of the independent variables. The poorest of the predictors were the first and second CAM test scores.

TABLE 3.4.2.4

Decisions for Prediction Equations with 1STUNIT as the Independent Variable for N = 256, N = 128 (Even-Numbered), and N = 128 (Odd-Numbered)

Predicted Grade (N)	Actual Grade: Pass		Fail	
	N	%	N	%
For N = 256				
Pass (218)	190	(87%)	28	(13%)
Fail (38)	11	(30%)	27	(70%)
	201	(79%)	55	(21%)
For N = 128 (even-numbered)				
Pass (112)	101	(90%)	11	(10%)
Fail (16)	4	(25%)	12	(75%)
	105	(82%)	23	(18%)
For N = 128 (odd-numbered)				
Pass (98)	85	(87%)	13	(13%)
Fail (30)	11	(37%)	19	(63%)
	96	(75%)	32	(25%)

3.4.2.5 Discussion. The decision-box approach proved useful in the prediction of failure although predicting success was more accurate as would be expected. A future development of the decision box may eventually enable a computer to "flag" performances of students as possible "fail" when a loss function identifies it as such. Such flagging would signal the need for special attention by a teacher or counselor with the student predicted as possible "fail."

One part of the analysis in need of revision was the arbitrary means of obtaining a decision-theoretic rule with which to test prediction possibilities. The arbitrary setting of a decision rule for probable pass-probable fail at the midpoint between "C" and "D" (i.e., between 3.0 and 2.0) had several problems:

(1) The composition of final grades was largely based on posttest performances and supposedly one-third of the final grade was for laboratory work which would not bear directly on performance on CAM and unit testing.

(2) There were "F" grades (i.e., 1.0) given. Thus, a grade of 2.0 was actually a passing grade.

(3) Grading was complicated by being done by four different teachers, two of whom attached "+" and "-" to the letter grades indicating at least one difference in grading practices.

(4) The choice of the final semester grade as a measure of end-of-course success is probably a poor choice. A "grade" is usually an average of grades given for performances over time. Thus it is not a reliable indicator of final success but one of success averaged across all performances. This, in turn, is complicated by the judgment of the grader which in the present case is actually four graders. Better criteria for success would be indicators of final performance on objectives. A suggestion is the last few CAM scores or other measures of final performances.

A virtue of using a cutoff score of 2.5 was to produce a conservative decision model for applying the prediction equations in simulation. Thus, both the "accuracy" of prediction (percent of those examinees predicted to pass or fail who actually passed or failed) and the "reliability" of prediction (percent of the predicted category which actually was correctly predicted) were low estimates of whom the teachers would have judged to be passing and high estimates of those failing. The surprising elements were the strengths of the SCAT score and first unit test as screening devices for course entry.

CHAPTER IV

CONCLUSIONS AND RECOMMENDATIONS

This chapter has three purposes:

(1) To provide a general summary of the results of the analyses for the six individual research questions and to present generalizations based on these individual results combined,

(2) To suggest research which would build on the conclusions of this study, and

(3) To point out specific applications of the results of this study for users of CAM and unit testing.

4.1 Summary of Results by Question

Question 1 investigated CAM test-form difficulties for differences at each test administration and for major differences in relative difficulty across administrations. The results are as follows:

(1) The eight CAM test forms differed in difficulty at five of the six CAM test administrations (all $p < .025$).

(2) All eight test forms became less difficult at each succeeding test administration.

(3) There was no significant change in relative difficulty of each test form to the others at each CAM test administration, i.e., the eight test forms maintained the same difficulty ranking across administrations.

(4) All eight student schedule groups were found to perform equivalently on the CAM tests and also on the unit tests.

A possible generalization from the results of Question 1 is that given student schedule groups which perform equivalently on the tests, test forms will not vary in relative difficulty to each other across test administrations. It would be of immense value to criterion-referenced testing to be able to make this statement particularly since teachers construct the vast majority of criterion-referenced tests. Also producing randomly parallel test forms requires field testing the items or, more practically for the teacher, trials in the actual instructional setting. The early products of these trials will be like the test forms in this study, unequal in difficulty. If there were assurance that forms would not fluctuate in relative difficulty, test results would be more reliable and useful to the teacher. However, this claim cannot be made unequivocally. Replication

of the study is required. Different subject areas and test form designs might not yield the same results.

Question 2 examined the effect on test-item difficulty of an item used in the two contexts of CAM and unit tests. A second comparison was made of differences in test-item difficulty of items where subjects had either previous or no previous experience with the item. The results are:

(1) There was no difference in item difficulty (i.e., student group performance) between test items with prior exposure to subjects versus items with no prior exposure.

(2) Items answered immediately after instruction on their associated objectives were found to be significantly more difficult if they appeared on CAM tests than if they appeared on unit tests.

No generalizations can be made from the results of Question 2 which are applicable to other populations. Two reasons for this are: (1) the format of the unit tests were similar to the CAM tests in that most objectives were measured by only one or two test items and (2) an adequate control for or measure of student motivation for taking either of the two types of tests was not present in the study.

Question 3 investigated the relationships among the global measures of course success and selected potential

predictors of these measures. The purpose was to identify possible predictor variables which have high correlations with the global measures of success (i.e., the criteria), but low correlations with each other. The results are:

(1) Early CAM test scores, the SCAT score, and the first unit test score all had rather high correlations with the global measures (most correlations were over .45 with nearly one-third over .60).

(2) The SCAT score and the first unit test score appeared to be better potential predictors of the global measures than early CAM scores.

The results obtained are certainly not of an unexpected nature. With cognitive measures, positive correlations are the rule. The results are also inclusive. There were many test items shared by both the predictor and criterion variables. Thus, the correlations produced are contaminated and to an unknown degree.

Question 4 focused on the usefulness of CAM scores in predicting global measures of individual student progress and the final semester grade. The practicality of using SCAT, CAM and unit test scores for prediction in a classroom situation was examined with simulated decision making with the result of each prediction equation being matched against final semester grade (the criterion). The results are:

(1) The first unit test score was the best predictor of final semester grade with the normalized SCAT score second best but not nearly as good. Early CAM scores (first and second CAM tests) were a distant third and fourth in usefulness for prediction.

(2) In the decision-making simulations, the first unit test score alone predicted 95% of the students designated as successes and 49% of those designated failures for N = 256. The normalized SCAT score alone predicted 96% of the successes but only 25% of the failures of N = 256. The addition of the first and/or second CAM test scores to the equations changed the percents negligibly.

These results bear out what might be conjectured as the case from only a surface knowledge of the situation. One would normally assume that high scores on a verbal ability test (the course of biology as it was taught required much reading and writing) and on early mastery tests in a course would be highly predictive of final success in the course. Another factor is the method used to measure success, i.e., in this case, the final semester teacher-assigned grade. The composition of the grade is important. The grade was a ranking of students based primarily on mastery test results and laboratory performance. The CAM tests were poor predictors of

this grade. To produce results with some generalizability about the value of CAM in prediction it probably would require repeating the design with other subject areas and grade levels.

One generalization which emerges is that given information about a student's ability relative to a course of study and early measures of his performance in mastering course objectives, the prediction of his eventual course success can be highly accurate.

4.1.1 Other conclusions related to the use of CAM.

This study made use of an actual classroom situation whereby the testing procedures were not superimposed on the normal instructional setting. Items were somewhat refined by teachers but an examination of item difficulties done preparatory to the analyses of this study showed most objectives to be represented by items which were not equivalent in difficulty. As the items for each objective become more equivalent, the tighter the individual CAM tests will be in terms of equivalence.

Emerging from the four analyses is the "robustness" of the CAM technique in delivering trend data against the odds imposed by situational factors. By "robustness" is meant the strength of the technique in furnishing information to teachers useful in everyday, practical decision making when certain testing parameters are less than perfect. For example, given the known nonequivalence of test items, the

analyses of Question 1 showed that the CAM tests did not significantly fluctuate in rank as to test-form difficulty across test administrations. Also the CAM technique for establishing student schedule groups was shown through the analysis of variance to have created no interaction effects attributable to student groups. The analysis of Question 4 showed CAM capable in the simulated decision-making analysis of having half the student population be as useful as the whole population in establishing prediction equations for student success.

Also apparent are the shortcomings of the longitudinal CAM design. It furnishes a conservative estimate of student performance which is probably due to a combination of the test design with only a single item sampled per objective and the relatively low motivation of students taking the tests compared to the motivation of taking unit posttests.

4.2 Recommendations for Future Research

This section opens with a discussion of the factors desirable for setting up a CAM design for a research study. This discussion is followed by a brief listing of ideas for future research designs which focus on the questions asked in this study.

4.2.1 Factors in designing a study with CAM. The following discussion is divided by the major variables of the CAM design.

(1) Teachers. A shortcoming of the present study was an unequal representation of students from each of the four teachers. The teachers differed in conscientiousness in using the CAM data and in having all students complete the testing treatment. There was also no means of systematically checking to see that each teacher was presenting material and maintaining techniques which were essentially equivalent to what the other teachers were doing.

A more thorough study would include more teachers so that individual difference in conscientiousness would be spread over more of a continuum and results become more generalizable.

(2) Subjects. The sample size of 256 students appeared adequate because the size of student schedule groups (SSGs) were sufficiently large (ranging 29 to 36) to insure reliable results from analyses of variance. However, one problem was the unequal size of SSGs. This left the largest group containing seven students more than the smallest. However, internal consistency of the groups (for CAM this is the representativeness of a given SSG) was excellent and a very necessary design component. A second problem was the lack of SCAT scores for all subjects. A third problem was the composition of the student body--a span of two school years, ninth and tenth grades. Tenth graders would probably have had another year of science and bring more of a background to the course than the ninth graders. However, no

means of substantiating differences were employed.

(3) Testing treatment. The duration of the treatment (one semester) appears to be long enough in that a large number of objectives and items can be accommodated on test forms and the span of 20 weeks allows up to 16 to 18 test administrations if one tests as frequently as weekly, but the trend with users of CAM is to test less frequently.

A second issue is the frequency of testing. CAM testings numbered six whereas if administrations had occurred biweekly, as is typical for a CAM system designed to measure performance on a large number of the instructional objectives, the number of administrations would have been from 8 to 10. This increase in the number of data points would have added to the value of the regression analyses. A basic limitation for the CAM longitudinal testing design appears to be the frequency with which a course representative test can be administered. Student motivation to respond to the tests would probably decrease as the frequency of testing increased. Variations on the item-sampling design need to be studied in relation to student motivation. Since student motivation is largely a product of the teacher's instructional methodology and attitudes the research could become quite complex.

A third issue is test content. In constructing the CAM test forms for this study, teachers used a rough rule as to the type of items put on the test forms. About 25% of

the objectives were selected to be measured by items on CAM tests identical to those on unit tests, another 25 % were selected to have some items identical on both types of tests and some different, and 50% were selected to have items on CAM tests different from the items on the unit tests. This composition admitted various choices for analysis of which only the identical items of both CAM and unit testing were used. The uncontrolled factors here were: (1) lack of adherence to the rule and (2) items having varying difficulty levels, i.e., they were not equivalent for a given objective. A more systematic method of establishing interchangeable test forms is needed.

A fourth issue was the extreme weighting of objectives and units. (See Table A-1 of Appendix A.) Objectives had from one to 10 items each, and were not always evenly divided between CAM tests and unit tests. Several items were also used twice on CAM tests. Review or course prerequisite objectives generally had only one or two items each which often appeared either on a CAM test or on the first unit test but not both. Units were weighted with from one to 16 objectives each and suffered the same problems of item representation on tests as did single objectives.

A fifth problem was the use of so few items for most objectives on both the CAM and unit tests. This was particularly deleterious on the unit tests because more than half of the objectives were tested by only one or two items on the unit tests providing a very limited demonstration by students

of knowledge of these objectives. Thus, the actual composition of the unit and CAM tests was highly similar in terms of weighting of objectives with test items. A larger difference in test composition would provide a better means of comparison and perhaps different results to the research questions of this study.

The test length (34 items for CAM forms and 30 and 40 items for unit tests) seemed appropriate for the age level of students particularly since items were multiple-choice items. However, multiple-choice items add fluctuation to individual scores because of the guessing factor involved versus the more time consuming and harder-to-correct, open-ended test items.

In sum, two recommendations for testing treatment stand out: (1) test content should be rigidly controlled to insure equivalent representation of individual objectives where possible, and (2) objectives should be represented by several items each on unit or mastery tests to provide stability and accuracy in measurement.

4.2.2 Future research designs. Research relating to criterion-referenced testing is in its infancy. The techniques used in this study need to be applied to other content areas and student populations to determine their general validity and scope of application. Suggestions for future research stemming from the questions of this present study follow:

1. Criteria for student schedule group (SSG) equivalence. Question 1 emphasized the importance of the equivalence (i.e., representativeness) of SSGs. The usual means of assigning students to SSGs are to randomly select students by selecting one student and placing him in SSG 1, another student and placing him in SSG 2, and so on until all students are placed in a SSG. A second means is to stratify students by some ability or achievement measures into three groups--low, medium, and high. Then selection is made randomly from each ability or achievement group to create the SSGs.

Variables such as student interest level and skill capability may suggest other dimensions of achievement which will affect student progress trends and result in learning curves valuable in program and curriculum analysis and validation. For instance, an interest level variable for establishing SSGs might produce data suggestive of instructional approach and sequence best suited to increasing interest in the course.

2. Effects of the two environments of CAM and unit testing. Question 2 determined that students perform differently on CAM and unit tests. This finding has implications for the analysis of results and the interpretation of test scores depending on the type of testing environment. More detailed approaches need to be devised than were used in this study to answer the question of whether or not stu-

dents respond differently in the two environments of CAM and unit testing.

One focus would be to study the CAM computer-generated item difficulties for individual items which are calculated for three instruction phases: preinstruction, immediate postinstruction, and long-term postinstruction retention. Comparisons can be made between these item difficulties and the postinstruction item difficulties of unit tests. More in-depth information would be available on the changing relationship between responses to items on CAM and unit tests.

3. Further study of the complementary use of CAM and and unit testing. Several questions on the complementary use of CAM and unit testing remain unanswered or only partially answered. A list of these follows:

(a) Can CAM tests be constructed so as to provide a subscore of posttest achievement predictive of unit achievement estimates?

(b) Are unit test scores predictive of CAM retention subscores? The question relates to the possibility of producing activities designed to prevent problems for content areas predicted to become areas of particular difficulty.

(c) Are CAM pretest scores predictive of unit scores? Tests constructed to allow prediction of posttest scores from pretest scores would be directly

useful in program refinement particularly in day-to-day organization of instruction.

4. Trend analysis of achievement. The longitudinal CAM testing design produces trend data. A major problem in the analysis of achievement patterns is that of establishing a sound procedure for determining the goodness of fit of group data based on various sizes of samples. A chi-square approach is one method, but other avenues bear investigation. In Question 2, nonparametric sign tests (the Wilcoxon and the Walsh tests) were used for a similar problem when other statistical approaches appeared not as satisfactory. The application of the nonparametric sign tests to goodness of fit in trend analysis needs to be investigated.

REFERENCES

Allen, D. W. First annual report to the Charles F. Kettering Foundation, Annual Report No. AR-1. School of Education, The University of Massachusetts, Amherst, 1968, Grant No. 642, C. F. Kettering Foundation.

Allen, D. W. & Gorth, W. P. Fourth annual report. Annual Report No. AR-4. School of Education, The University of Massachusetts, Amherst, 1971, Grant No. 642, C. F. Kettering Foundation.

Atkinson, R. C. Computer-based instruction in initial reading. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1968.

Atkinson, R. C., & Wilson, H. A. (Eds.) Computer-assisted instruction: A book of readings. New York: Academic Press, 1969.

Bloom, B. S. (ed.) Taxonomy of educational objectives: The classification of educational goals. Handbook 1. Cognitive domain. New York: McKay, 1958.

Bloom, B. S., Hastings, J. T. & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Dodd, S. C. Operational definitions operationally defined. American Journal of Sociology, 1943, 48, 482-489.

Flanagan, J. C. Functional education for the seventies. Phi Delta Kappan, 1967, 49, 27-32.

Flanagan, J. D. Program for learning in accordance with needs. Psychology in the Schools, 1969, 6, 133-136.

Gagne, R. M. The conditions of learning. New York: Holt, Rinehart and Winston, 1965.

Glaser, R. Adapting the elementary school curriculum to individual performance. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1968, 18, 519-521.

Glaser, R. & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement, second edition. Washington: American Council on Education, 1971, pp. 625-670.

Gorth, W. P., Allen, D., Popejoy, L. & Stroud, T. The relation of repeated comprehensive pretesting to student achievement. Technical Memorandum No. TM-4. School of Education, The University of Massachusetts, Amherst, 1968, Grant No. 642, C. F. Kettering Foundation.

Gorth, W. P. & Grayson, A. Computer programs for objective and item banking. Educational and Psychological Measurement, 1971, 31, 245-250.

Gorth, W. P., Schriber, P. E., & O'Reilly, R. P. Comprehensive achievement monitoring. Amherst, Massachusetts: School of Education, The University of Massachusetts, 1971.

Gorth, W. P., Wightman, L., O'Reilly, R. P. & Schriber, P. E. CAM and the computer help teachers improve evaluation. Trend, 1970, Spring, 5-6.

Hambleton, R. K. Item analysis program. Unpublished computer program. School of Education, The University of Massachusetts, 1970.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method of criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harvey, W. R. Least-squares and maximum likelihood general purpose program. Unpublished manuscript on use of a computer program. Ohio State University, 1968.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. Taxonomy of educational objectives: The classification of educational goals. Handbook 2. Affective domain. New York: McKay, 1964.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Mager, R. F. Goal analysis. Belmont, California: Fearon, 1972.

Mager, R. F. Preparing objectives for programmed instruction. Belmont, California: Fearon, 1962.

Mayo, S. T. Book review of James W. Popham (Ed.) Criterion-referenced measurement. Psychometrika, 1972, 37, 489-490.

McCall, W. A. How to experiment in education. New York: Macmillan, 1923.

Millman, J. Criterion referenced measurement: An alternative. Reading Teacher, December 1972, 278-281. (a)

Millman, J. Passing scores and test lengths for domain-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972 (ERIC ED 065 555). (b)

Millman, J. Tables for determining number of items needed on domain-referenced tests and number of students to be tested. Technical Paper No. 6, Los Angeles: UCLA, Instructional Objectives Exchange, April, 1972. (c)

Myers, J. L. Fundamentals of experimental design. Boston: Allyn and Bacon, 1966.

O'Reilly, R. P. The conceptualization of objectives for evaluation. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

O'Reilly, R. P. & Gorth, W. P. Alternatives to accountability: Stool pigeon versus servant and soulmate. Working Paper No. WP-31. School of Education, The University of Massachusetts, Amherst, 1971. (Published 1972).

O'Reilly, R. P., Gorth, W. P. & Pinsky, P. Computer assisted test construction: An effort based on an evaluation methodology. Educational Technology, March 1973, 13, 32-34.

Osburn, H. G. The effect of item stratification on errors of measurement. Educational and Psychological Measurement, 1969, 29, 295-301.

Pinsky, P. D. A Comprehensive Achievement Monitoring (CAM) data processing system. Technical Memorandum No. TM-32. School of Education, The University of Massachusetts, Amherst, 1971, Grant No. 642, C. F. Kettering Foundation.

Pinsky, P. D. & Gorth, W. P. Descriptive analysis of HS420: Eleventh Grade algebra first semester. Technical Memorandum No. TM-21. School of Education, The University of Massachusetts, Amherst, 1969, Grant No. 642, C. F. Kettering Foundation (ERIC Ed 042 794). (a)

Pinsky, P. D. & Gorth, W. P. Descriptive analysis of KA442: One Semester eleventh and twelfth grade trigonometry. Technical Memorandum No. TM-22, School of Education, The University of Massachusetts, Amherst, 1969, Grant No. 642, C. F. Kettering Foundation (ERIC Ed 042 796). (b)

Popham, J. W. (Ed.) Criterion-referenced measurement. Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

Popham, J. W. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Education Measurement, 1969, 6, 1-9.

Schriber, P. E. & Gorth, W. P. Objective and item banking computer software and its use in Comprehensive Achievement Monitoring. Working Paper No. WP-17. A paper presented at the Annual Meeting of the American Educational Research Association, New York, February, 1971, Grant No. 642, C. F. Kettering Foundation.

Shoemaker, D. M. Allocation of items and examinees in establishing a norm distribution by item-sampling. Journal of Educational Measurement, 1970, 7, 123-128. (a)

Shoemaker, D. M. Item-examinee sampling procedures and associated standard errors in estimating test parameters. Journal of Educational Measurement, 1970, 7, 255-262. (b)

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Sund, R. B. & Picard, A. J. Behavioral objectives and evaluational measures, Columbus, Ohio: Charles E. Merrill, 1972.

Suppes, S. The uses of computers in education. Scientific American, 1966, 215, 206-221.

Swaminathan, H. Statistical procedures for the analysis of longitudinal data obtained by matrix sampling. Final Report. Albany, New York: Bureau of School and Cultural Research, New York State Education Department, 1972.

Tanner, D. Using behavioral objectives in the classroom. New York: Macmillan, 1972.

UCLA - Health Sciences Computing Facility,

(a) BMD01V Analysis of variance for one-way design - version of June 11, 1964.

(b) BMD03R Multiple regression with case combinations - version of August 13, 1964.

(c) BMD05R Polynomial regression - version of August 15, 1964.

(d) BMD02R Stepwise regression - version of June 2, 1964.

In W. J. Dixon (ed.) BMD biomedical computer programs, second edition. Los Angeles, University of California Press, 1967.

Vargas, J. S. Writing worthwhile behavioral objectives. New York: Harper & Row, 1972.

APPENDIX A

Table Showing Test Form Composition for CAM Tests and Unit Tests

This appendix contains a tabular arrangement of the test-form composition of both the CAM tests and unit tests used in this study. It is a breakdown by item of each test form and contains each item, the objective to which the item is associated, the units containing the objective, the lesson containing the unit, and the data of lesson completion. The key for the table (Table A-1) is as follows:

(1) CAM test form numbers are given across the top (Test Forms 51 to 58).

(2) Unit test form numbers are given in the right half of the table immediately above the items contained on the form.

(3) At the far left margin:

(a) Time: refers to date of lesson completion,

(b) Lesson: gives the lesson number designation for each objective whose identification number is to the immediate right, and

(c) Obj(ective): designates the objective identification number, the first two digits of which are also the unit identification number.

(4) The columns headed by CAM test form numbers contain the item identification numbers and position of these items on the CAM test forms. The identification numbers are read as follows, under CAM Test Form 51 is "01(8)" which means that Item 110301 of Objective 1103 is in the 8th position on the test form. Several objectives have two items on a given test form and thus two pairs of item identification numbers and position numbers are set side-by-side without spacing between them in the appropriate column.

(5) Items for unit test forms are represented the same way with the exception that the forms are represented horizontally in the table with one above the other rather than vertically like the CAM forms which are presented side-by-side.

(6) The total number of items per CAM form is given at the bottom of the page.

TABLE A-1

Test Items by Test Form, Objective, Unit, Lesson, and Date of Lesson Completion

Time	Lesson	Obj	CAM Form Number Under Which Are Items (by Position on Form) 51	52	53	54	55	56	57	58	Unit Test Forms Under Which Are Items (by Position on Form)
											Unit Test Form 61 - 40 Items
9/20	1	1103	01(8)								01(1)
	1	1104			01(27)						06(2)
	1	1201									01(3)
	1	1202									03(4)
	1	1204									02(5)
	1	1205					01(8)				01(6)
	1	1206								01(8)	02(7)
	1	1207		01(8)							01(8)
	1	1301				01(8)		02(8)			03(9)01(10)
	1	1302			04(8)				03(8)		03(11)04(12)
	1	1303		03(27)		01(27)					03(13)05(14)09(15)
	1	1304					02(2)01(27)				02(16)
	1	1305	01(27)							01(27)	01(17)
	1	1306									04(18)
	1	1307						02(27)	01(27)		01(19)02(20)
10/1	2	1402	01(4)		02(2)						05(21)06(40)
	2	1501		01(4)					01(2)		01(22)
	2	1502						03(4)		01(2)	02(23)
	2	1601		01(2)	02(4)						01(24)03(25)
	2	1602				02(4)	03(4)				03(26)05(27)
	2	1603	01(2)			04(2)			05(4)		04(28)02(29)
	2	1604						01(2)		05(4)	02(30)
10/13	3	2101	01(26)	09(9)08(26)	07(26)	06(9)	05(26)	04(26)	03(9)	02(26)	01(31)02(32)05(33)09(34)06(35)
	3	2102									02(36)
	3	2103	01(9)			02(26)					03(37)
	3	2104			03(9)			04(9)		05(9)	01(33)
	3	2105					01(9)		02(26)		05(39)

TABLE A-1 (Continued)

											Unit Test Form 62 - 40 Items
											02(1)
10/20	4	2201	01(3)			03(34		02(34)		01(34)	02(2)
	4	2203		06(34)		04(3)			05(3)		03(3)
	4	2204	01(34)		02(3)						
	4	2205		02(3)			01(3)03(34)	04(3)	05(34)	06(3)	01(4)
	4	2206			01(34)						05(5)
10/25	5	2301	01(23)		08(23)		03(23)		02(23)		02(6)
	5	2302	01(21)	05(21)				02(23)		03(21)	04(7)
	5	2303		02(23)		03(21)01(23)		04(21)		05(23)	09(8)
	5	2304			01(21)				02(21)		01(9)
							02(21)	03(6)	04(6)		03(10)
10/28	6	2305	01(6)	05(16)			02(6)		03(16)	04(16)	04(11)
	6	2306				01(16)		02(16)			01(12)02(13)
	6	2307			01(16)						
										05(6)	06(14)
	6	2308	01(16)	02(6)	03(6)	04(6)	06(16)				02(15)01(16)
11/23	7	2401	01(19)				02(19)		01(19)		04(17)
	7	2402			02(19)	03(19).	04(5)				
											04(18)
	7	2403		02(5)01(19)	03(5)	04(5)		05(5)	01(5)	02(19)	02(19)
	7	2404								04(5)	02(20)
	7	2405	01(5)					03(19)			
									10(15)	20(15)	08(21)15(22)22(23)
11/23	8	2501	01(12)	02(12)	05(12)	07(15)					03(24)01(25)
	8	2502	01(15)		06(15)	07(12)	03(15)		01(12)	02(12)	03(26)01(27)
	8	2503									
							04(12)	01(12)			01(28)
	8	2504									05(29)
	8	2505		04(15)					04(32)	02(32)	06(30)
11/23	9	2506	01(32)								
								02(7)			01(31)
	9	2507		01(7)	07(7)	06(7)		05(15)06(32)		07(7)	04(32)
	9	2508	01(7)		02(32)		03(32)				04(33)
	9	2509		02(32)							
									01(7)		11(34)
	9	2510				05(32)	04(7)		04(11)		07(35)
11/23	10	2511	01(30)		03(11)		02(11)			01(30)	01(36)
	10	2512		02(11)							
									09(30)		03(37)
	10	2513	03(11)		06(30)	05(11)		01(30)			02(38)
	10	2514									05(39)
	10	2515				01(30)				03(11)	03(40)
	10	2516		04(30)			01(30)	02(11)			

TABLE A-1 (Concluded)

											Unit Test Form 63 - 40 Items
12/7	11	2601	01(31)		03(13)	10(13)					01(1)04(2)05(3)07(4)08(5)09(6)10(7)
	11	2602			04(31)		06(13)		12(13)		01(8)02(9)04(10)07(11)09(13)14(15)
	11	2603				03(31)				04(31)	03(12)14(14)
	11	2604		02(13)			01(31)				01(16)02(17)
	11	2605						01(13)			01(18)02(19)
	11	2606							01(31)	02(13)	03(20)
	11	2607	05(13)	02(31)				03(31)			05(21)02(22)
	11	2608									02(23)03(24)
12/10	12	2701	01(25)	02(25)	03(18)	01(25)	02(18)		03(25)		03(25)
	12	2702		01(18)						02(18)	02(26)
	12	2703	01(18)		08(25)	07(18)	05(25)	06(18)04(25)	03(18)	02(25)	01(27)03(28)02(29)06(30)
12/17	13	2704	01(28)	07(28)	04(22)03(28)		02(22)		05(22)		04(31)09(32)
	13	2706	01(22)								01(33)02(34)03(35)
	13	2707		07(22)		08(22)01(28)	05(20)	04(22)03(28)	02(28)	09(22)06(28)	01(36)03(37)04(38)09(39)10(40)
											Unit Test Form 64 - 30 Items
											01(1)04(2)05(3)
1/11	14	2801	01(29)			03(1)		02(29)			02(4)
	14	2802	01(1)		02(1)				02(1)	01(1)	01(5)
	14	2803									
	14	2804		05(1)04(29)		83(29)	02(1)	01(1)	01(29)		04(6)05(7)03(8)
	14	2805			09(29)		06(29)			01(29)	01(9)02(10)
1/28	15	2901	01(10)				02(10)				01(11)
	15	2902		09(14)	13(14)	14(14)				15(14)	13(12(15(13)
	15	2903					13(14)	05(14)	02(14)		
	15	2904	01(14)	06(10)					05(10)	02(10)	08(14)10(15)12(16)19(17)23(18)26(19)
	15	2905			10(10)	08(10)		01(10)			01(20)10(21)
1/28	16	3001	02(20)	14(17)	13(20)	17(17)09(20)	07(20)	06(20)	05(17)04(20)	03(20)	01(22)05(22)08(24)09(25)14(26)16(27)17(28)
	16	3002	01(17)	07(20)	06(17)		04(17)	03(17)		02(17)	03(29)
										02(33)	
1/28	17	3003	01(33)		05(33)	04(33)		03(33)		01(24)	01(30)
	17	3004	01(24)	07(24)05(33)	06(24)	02(24)	04(24)03(33)	06(24)	02(24)05(33)		
	Total		34	34	34	34	34	34	34	34	

APPENDIX B

Tables of the Item Difficulty of 10 Items of Unit 26 Which Appear on Both CAM and Unit Test Forms by Student Schedule Group (SSG) and by Test Administration

In the 10 tables of this appendix the type of test and the test form identification number are given immediately below the test administration number.

TABLE B-1

Item Difficulty of Item 260101 by SSG and by Test Administration

SSG (and Size)	Test Administration 1 CAM 51	2 CAM 51	4 CAM 51	6 CAM 51	7 Unit 63	8 CAM 51	9 CAM 51
1(29)	.21				.83		
2(30)			.23		.83		
3(36)					.86		
4(33)				.56	.88		
5(31)					.87	.71	
6(35)		.31			.94		
7(33)					.88		.79
8(29)					.82		

TABLE B-2

Item Difficulty of Item 260110 by SSG and by Test Administration

SSG (and Size)	Test Administration 1 CAM 54	2 CAM 54	4 CAM 54	6 CAM 54	7 Unit 63	8 CAM 54	9 CAM 54
1(29)					.83		
2(30)				.71	.87		
3(36)					.92		1.00
4(33)	.53				.67		
5(31)		.26			.87		
6(35)					.91		
7(33)			.48		.82		
8(29)					.86	.83	

TABLE B-3

Item Difficulty of Item 260204 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 53	2 CAM 53	4 CAM 53	6 CAM 53	7 Unit 63	8 CAM 53	9 CAM 53
1(29)					.69		
2(30)					.70	.73	
3(36)	.24				.84		
4(33)		.24			.76		
5(31)			.23		.87		
6(35)				.83	.91		
7(33)					.79		
8(29)					.68		.90

TABLE B-4

Item Difficulty of Item 260303 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 54	2 CAM 54	4 CAM 54	6 CAM 54	7 Unit 63	8 CAM 54	9 CAM 54
1(29)					.62		
2(30)				.61	.73		
3(36)					.81		.61
4(33)	.18				.70		
5(31)		.13			.77		
6(35)					.83		
7(33)			.27		.70		
8(29)					.89	.77	

TABLE B-5

Item Difficulty of Item 260304 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 58	2 CAM 58	4 CAM 58	6 CAM 58	7 Unit 63	8 CAM 58	9 CAM 58
1(29)				.76	.72		
2(30)					.97		.83
3(36)		.59			.84		
4(33)			.56		.94		
5(31)					.83		
6(35)					.80	.74	
7(33)					.70		
8(29)	.52				.75		

TABLE B-6

Item Difficulty of Item 260401 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 55	2 CAM 55	4 CAM 55	6 CAM 55	7 Unit 63	8 CAM 55	9 CAM 55
1(29)			.40		.48		
2(30)					.67		
3(36)					.81	.69	
4(33)					.67		.74
5(31)	.32				.67		
6(35)					.80		
7(33)		.33			.79		
8(29)				.52	.75		

TABLE B-7

Item Difficulty of Item 260402 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 52	2 CAM 52	4 CAM 52	6 CAM 52	7 Unit 63	8 CAM 52	9 CAM 52
1(29)					.55	.55	
2(30)	.20				.60		
3(36)			.38		.62		
4(33)					.58		
5(31)					.57		
6(35)					.63		.60
7(33)				.42	.52		
8(29)		.17			.71		

TABLE B-8

Item Difficulty of Item 260501 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 56	2 CAM 56	4 CAM 56	6 CAM 56	7 Unit 63	8 CAM 56	9 CAM 56
1(29)					.90		.69
2(30)		.23			.80		
3(36)					.81		
4(33)					.73		
5(31)				.61	.73		
6(35)	.37				.69		
7(33)					.73	.70	
8(29)			.24		.79		

TABLE B-9

Item Difficulty of Item 260702 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 52	2 CAM 52	4 CAM 52	6 CAM 52	7 Unit 63	8 CAM 52	9 CAM 52
1(29)					.62	.66	
2(30)	.37				.70		
3(36)			.49		.65		
4(33)					.67		
5(31)					.87		
6(35)					.60		.60
7(33)				.67	.57		
8(29)		.52			.68		

TABLE B-10

Item Difficulty of Item 260705 by SSG
and by Test Administration

SSG (and Size)	Test Administration 1 CAM 51	2 CAM 51	4 CAM 51	6 CAM 51	7 Unit 63	8 CAM 51	9 CAM 51
1(29)	.10				.41		
2(30)			.27		.47		
3(36)					.35		
4(33)				.32	.30		
5(31)					.20	.42	
6(35)		.14			.29		
7(33)					.52		.52
8(29)					.25		

APPENDIX C

Summary of the Research and Thought on Criterion-Referenced Testing Extending the Background (Section 1.1) of Chapter I through 1974

Glaser and Nitko (1971) defined a criterion-referenced test as "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). While this definition was more specific than previous definitions in that it stated "deliberately constructed" and "directly interpretable," it, nonetheless, covers a broad category and makes no differentiation between tests composed of items which were systematically generated from a precise set of item generation rules or of items judged by a single item reviewer as having face validity and probable reliability. Perhaps due to a general, pervasive vagueness in the definition and, hence, test construction procedures, the concept of criterion-referenced measurement has been used more in the interpretation of test results than in the methodology of a testing procedures.

Millman (1974) attempts to clarify what has heretofore been termed criterion-referenced tests. He focuses on the specificity of test content, i.e., the specificity of the description of the item population. Both Millman (1974) and Hively (1974) prefer the term "domain-referenced tests" (DRT) rather than "criterion-referenced tests." The word "criterion" is perhaps suggestive of a criterion test, a term which denotes a special use of a test in an experimental design. The word "domain" suggests that an item population is defined. Millman states that the domain "may be extensive or a single, narrow objective, but it must be well defined, which means that con-

tent and format limits must be well specified." (p. 314). Thus, he offers the definition for a DRT as a test composed of items (each constructed with a high degree of specificity to measure performance of an objective) which are drawn from a population of such items in a random or stratified random fashion. Therefore, test content is clearly described and test scores can be directly interpretable in terms of the objectives (i.e., performance standards).

Additionally, a new clarity is introduced to making criterion-referenced interpretations. The random or stratified random sampling from a well-defined item population (i.e., the domain) permits "an estimation of an examinee's domain score or level of functioning, defined as the percent of the population of items the examinee could answer correctly or in a given direction" (p. 315).

To further crystalize the concept of DRT it is fruitful to contrast it with norm-referenced testing. These approaches to testing differ not only in the comprehensiveness with which the item population is specified but also in terms of how the items are constructed. Norm-referenced tests are constructed of items which differentiate among examinees on an attribute which the test is designed to measure. Millman (1974) terms such tests differential assessment devices (DAD's). DRT items are those which are related directly to the measurement of a specific content domain (for instance, an instructional objective). DAD items must have item characteristics

which differentiate among examinees on a particular attribute. Therefore, in general, a test cannot provide results directly interpretable in terms of performance tasks and also be optimal for differential assessment. The items necessary for both tests are drawn from different item pools. This situation has not been adequately focused on in the past in the literature. The echo of Popham and Husek's (1969) statement that a test may be used either as a norm-referenced or criterion-referenced instrument and only the interpretation of the results need distinguish the use is, at the very least, an oversimplification of the concept of criterion-referenced measurement.

The definition of a domain or the procedures for the construction of a domain of items is currently the eager concern of much research and many leading people in the field. Alkin (1974) has contrasted a few of the major researchers work in terms of the burgeoning jargon which is entering the field of criterion-referenced measurement. Apparently nearly every major contributor is coining a series of terms to describe his approaches. Of particular interest is the work of Baker (1974). She points up the inadequacy of behavioral objectives as a self-sufficient basis for generating test items. She identifies five necessary components: (1) the behavioral objective ("domain description"), (2) rules for determining content for the test items ("content limits"), (3) rules for judging the adequacy of responses to items, (4) item format, and (5) test directions. Procedures for item generation for

two nationally known endeavors — the Instructional Objectives Exchange (Popham, 1974) and the National Assessment of Educational Progress (Wilson, 1974) — provide practical situations in which varieties of item generation procedures for specific applications were designed and carried out.

A word should be said about the types of criterion-referenced tests being constructed. Items may be sampled from well-defined content domains but they may not all be placed on a single test form. They may be spread across test forms for administration of many items from the domain to subgroups of examinees sampled from the population. Thus, item-examinee sampling, a form a multiple matrix sampling, provides a large pool of items to be used over a given sample of examinees. A good discussion of multiple matrix sampling and of procedures in its use are given by Shoemaker (1973). A well-documented application of item-examinee sampling with criterion-referenced tests as part of a well-integrated criterion-referenced evaluation system is presented by Gorth, O'Reilly, and Pinsky (1974) in their description of the Comprehensive Achievement Monitoring model.

NEW REFERENCES FOR APPENDIX C

Alkin, M. C. "Criterion-referenced measurement" and other such terms. In C. W. Harris, M. C. Alkins, & W. J. Popham (eds.). Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974, pp. 3-12.

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-16.

Gorth, W. P., O'Reilly, R. P., & Pinsky, P. D. Comprehensive Achievement Monitoring: A criterion-referenced evaluation system. Amherst, Mass.: The CO-OP, Center for Educational Research, University of Massachusetts, 1974.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

Millman, J. Criterion-referenced measurement. In W. J. Popham (ed.) Evaluation in education: Current applications. Berkeley, California: McCutchan, 1974, pp. 309-398.

Popham, W. J. Selecting objectives and generating test items for objectives-based tests. In C. W. Harris, M. C. Alkins, & W. J. Popham (eds.). Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974, pp. 13-25.

Shoemaker, D. M. Principles and procedures of multiple matrix sampling. Cambridge, Mass.: Ballinger, 1973.

Wilson, H. A. A judgmental approach to criterion-referenced testing. In C. W. Harris, M. C. Alkins, & W. J. Popham (eds.). Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974, pp. 26-36.